Published in final edited form as:
Proteins. 2013 May ; 81(5): 805–818. doi:10.1002/prot.24235.

Predicting Permanent and Transient Protein-Protein Interfaces

David La[1,3], **Misun Kong**[1], **William Hoffman**[1], **Youn Im Choi**[1,3], and **Daisuke Kihara**[1,2,3,*]

[1]Department of Biological Sciences, College of Science, Purdue University, West Lafayette, IN, 47907, USA

[2]Department of Computer Science, College of Science, Purdue University, West Lafayette, IN, 47907, USA

[3]Markey Center for Structural Biology, Purdue University, West Lafayette, IN, 47907, USA

Abstract

Protein-protein interactions are involved in many diverse functions in a cell. To optimize functional roles of interactions, proteins interact with a spectrum of binding affinities. Interactions are conventionally classified into permanent and transient, where the former denotes tight binding between proteins that result in strong complexes, while the latter compose of relatively weak interactions that can dissociate after binding to regulate functional activity at specific time point. Knowing the type of interactions has significant implications for understanding the nature and function of protein-protein interactions. In this study, we constructed amino acid substitution models that capture mutation patterns at permanent and transient type of protein interfaces, which were found to be different with statistical significance. Using the substitution models, we developed a novel computational method that predicts permanent and transient protein binding interfaces in protein surfaces. Without knowledge of the interacting partner, the method employs a single query protein structure and a multiple sequence alignment of the sequence family. Using a large dataset of permanent and transient proteins, we show that our method performs very well in protein interface classification. A very high Area Under the Curve (AUC) value of 0.957 was observed when predicted protein binding sites were classified. Remarkably, near prefect accuracy was achieved with an AUC of 0.991 when actual binding sites were classified. The developed method will be also useful for protein design of permanent and transient protein binding interfaces.

Keywords

Protein-protein interaction; protein binding interface; protein-protein interaction network; permanent and transient interactions; phylogenetic substitution model; mutation pattern; sequence analysis

[*]Corresponding Author: dkihara@purdue.edu, Tel: (765) 496-2284, Fax: (765) 496-1189.

Introduction

Protein-protein interactions (PPI) mediate many essential functions of the cell [1-4]. Proteins interact with each other with different affinities for specific functional reasons. For example, enzyme-inhibitor, antigen-antibody, and large oligomeric enzyme complex structures are composed of proteins that are required to bind tightly and permanently (permanent interaction). In contrast, some proteins involved in signaling pathways [5-8] have a mechanism for dissociation after binding, thus help regulate protein activity at specific times (transient interaction). In a recent review, it was estimated that transient interactions make up a significant portion of protein-protein interaction networks [9-11]. Distinguishing between the two interaction types provide clues for functions of interacting proteins and have important implications for furthering the understanding the functional diversity exhibited in protein-protein interaction networks. Permanent and transient interactions are distinguished by the strength of the interactions, thus, in principle, can be associated with the dissociation constant (K_d) as long as two proteins dissociate. Dissociation constants of strongly permanent complexes are typically determined to be in the *nM* range (1×10^{-9} *M*) or lower[12-14], while transient complexes commonly show K_d in the *μM* range or higher (1×10^{-6} *M*)[15-17]. Further detailed discussion of this common range of dissociation constants between permanent and transient protein-protein interactions can be found in the recent reviews[18, 19].

There are several studies conducted in the past to understand differences in permanent and transient protein interaction sites[9, 19-21]. Permanent interaction sites are more conserved than transient interfaces and tend to have more hydrophobic residues, while transient interfaces consist of more polar residues[22-24]. Further, permanent interfaces tend to have fewer gaps in multiple sequence alignments of protein families than transient interfaces [25]. In terms of the size of protein interfaces, transient complexes form smaller interfaces than permanent interfaces[23]. Using physical and chemical properties, classification of crystal structures of bound protein complexes into permanent or transient types was attempted[24, 26, 27]. These methods require a known structural complex and thus their applications are limited to protein complexes whose structures have been experimentally determined.

In this work, we introduce a new method that predicts permanent and transient protein-protein interaction sites on a protein surface. The difference between the aforementioned existing methods and our method is that while the existing methods need an experimentally solved structure of protein complex and use actual interface residues as input data, our method uses the single protein structure only and make predictions of permanent and transient interface regions on the protein surface. Thus, prior knowledge of the interacting proteins nor interface regions is not needed. To the best of our knowledge, our method is the first of its kind.

In our previous work, we developed a computational method, called BindML, for predicting the protein binding interface (PBI) of a protein with unknown interacting partners[18]. BindML estimates the likelihood that a phylogenetic tree of a local surface region follows the amino acid substitution patterns of PBIs and non-PBIs. Through a comprehensive benchmark across a diverse set of protein structures, BindML was shown to perform better compared to alternative state-of-the-art methods, ProMate[28] and cons-PPISP [29], that combine various sequence and structural information into machine learning frameworks, for protein binding interface prediction.

Here, we extend the BindML framework for investigating mutation patterns of permanent and transient PBI sites. We built transient-PBI and permanent-PBI specific substitution models from a large dataset of permanent and transient protein complexes. Subsequently, we developed a multi-step procedure that is based on the BindML framework for predicting permanent and transient PBIs in a given protein surface. The method first predicts PBI sites in a protein surface, followed by discrimination of the sites either to permanent or transient PBI sites. It was shown that our method was able to distinguish between permanent or transient sites of PBI predictions with a very high accuracy of an Area Under the Curve (AUC) of 0.957. Moreover, the accuracy in classification of permanent and transient PBIs was raised to an AUC of 0.991 when actual PBI sites were provided. Overall, it was shown that the different functional requirements for permanent and transient PBI sites reflect to their mutation patterns, which can be used for distinguishing them. The method, named BindML+, is made available for the academic research community as a web server at http://kiharalab.org/bindml/plus/.

Methods

Dataset of Permanent and Transient Protein Complexes

Known permanent and transient protein complexes were used to construct amino acid substitution models and to benchmark the performance of protein-protein interface predictions and classifications. We initially considered 90 permanent interacting protein structures (from 39 permanent complexes) [30] and 145 transient protein structures (from 45 complexes), in which we designate as the Jones, Nooren, and Thornton (JNT) dataset[31]. Further, we added 161 permanent structures (71 permanent complexes) defined as those with dissociation constant (K_d) values of 1.0×10^{-9} or lower and 78 transient structures (33 transient complexes) as those with weak K_d values of 1.0×10^{-6} and higher from the Affinities dataset[1]. 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 value ranges and the associated number of complexes in the Affinity dataset is shown in Table I. Protein complexes with a Kd value between 10^{-9} and 10^{-6} are not used because there is no cutoff value that clearly distinguishes permanent and transient complexes, because classification between the two classes has been often done by considering other information, such as the functions of the complexes (e.g. pathways the proteins belong to).

Next, we combined the JNT and Affinities dataset together and removed redundant structures with 30% sequence identity, proteins that are annotated as monomers by PISA[5], and proteins that do not have PFAM[9] assignments in the dataset. Our final accumulative non-redundant dataset contains 110 permanent and 72 transient structures shown in Tables S1 and S2 in the supplementary materials.

Computing Interface Specific Substitution Models

We constructed amino acid substitution models (matrices) of permanent and transient PBI (termed PERM and TRAN, respectively) and non-PBI (non-permanent, NPERM; and non-transient, NTRAN). The amino acid substitution models reflect the ratio of pairwise amino acid substitutions observed in multiple sequence alignments (MSAs) of proteins against amino acid pairs appearing by chance. We use the same procedure as we described in our previous work[18], which is described as follows.

Given each protein structure in our permanent and transient protein dataset, we used the full set of sequences taken from the PFAM database[9]. Sequences from PFAM and the query sequence from the PDB structure [2-4, 32] were then used to construct a MSA using MUSCLE[6-8, 33]. 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[10, 11, 34]. Among the protein surface, residues at the interface (either permanent or transient) were defined as those that are closer than 5.0 Å to any residues in the protein docking partner, otherwise residues were defined as non-binding interface (NPERM or NTRAN). The observed substitutions were counted at gapless positions in the set of pairwise set of alignments following the JTT procedure[19, 35]. The values in the substitution matrices were calculated using the BLOSUM method[19-21, 36]. The resulting log odds matrices generated for PERM, NPERM, TRAN, and NTRAN are given in Table II.

Algorithm for Predicting Permanent and Transient Protein Interaction Sites

Our algorithm, BindML+, predicts permanent and transient protein interfaces of a query protein in two steps. The first step is to identify PBI in the protein surface. Subsequently, the PBI is classified to either permanent or transient interface. Figure 1 illustrates the whole procedure.

The first step identifies potential PBI sites employing the BindML algorithm developed in our previous work[18, 22] (Upper half of Figure 1). Input data are a PDB structure of a query protein and a MSA of its family including the query sequence. For each surface residue, a surface patch is defined as neighboring residues that are within the sphere of 15Å. The β-carbon of a given amino acid (α-carbon is used for glycine) is selected as the representative point when computing the distance between amino acids. For a patch, columns in the MSA that correspond to residues in the patch are concatenated together to form a "mini" MSA (patch MSA).

Then, a modified PHYML program is used to compute the likelihood that a patch MSA comes from PBI (involving permanent or transient binding sites) and non-PBI by constructing phylogenetic trees with either of the substitution models for PBI residues (PERM or TRAN) or NPBI residues (NPERM or NTRAN). PHYML computes the likelihood of having the input patch MSA and a tree topology using the PBI/NPBI substitution model. Finally, the difference of the likelihood under the PBI and NPBI substitution models provides a score that a center residue in the patch belongs to PBI. More concretely, for a patch MSA, P_i, which has residue i at the center, the log likelihood that the center residue i of the patch P_i is at non-PBI is

$$\mathrm{L_{NPBI}}\left(\mathrm{i}\right)=\log\left\{\mathrm{Prob}\left(\mathrm{P_i},\ \mathrm{T_{NPBI}}\left(\mathrm{i}\right)\ \middle|\ \mathrm{M_{NPBI}}\right)\right\} \quad (1)$$

Similarly, the likelihood that the center residue *i* of the patch P_i is at PBI is

$$\mathrm{L_{PBI}}\left(\mathrm{i}\right)=\log\left\{\mathrm{Prob}\left(\mathrm{P_i},\ \mathrm{T_{PBI}}\left(\mathrm{i}\right)\middle|\mathrm{M_{PBI}}\right)\right\}, \quad (2)$$

where M_{NPBI} and M_{PBI} is the substitution model of NPBI and PBI, respectively, and T_{NPBI}(i) and T_{PBI}(i) are trees generated with M_{NPBI} and M_{PBI}, respectively, for the input patch MSA that has residue *i* at the center. Note that T_{NPBI}(i) and T_{PBI}(i) are not necessarily identical.

Finally, the difference between the log likelihood of the patch MSA being NPBI and PBI, the distance likelihood (*dL*) score, is used for prediction:

$$\mathrm{dL}\left(\mathrm{i}\right)=\mathrm{L_{NPBI}}\left(\mathrm{i}\right)-\mathrm{L_{PBI}}\left(\mathrm{i}\right), \quad (3)$$

Once the *dL* scores for all surface patches in the query protein are computed, these scores are recast into *Z*-scores and a threshold is placed. A lower (negative) *Z*-score indicates larger likelihood of PBI mutation patterns, while a higher *Z*-score corresponds to less likelihood of following the PBI substitution model. Any center residue of a patch with a *dL* score that is equal to or smaller than a given *Z*-score threshold value is predicted to be included in a PBI site.

As the PBI and NPBI substitution models, we used PERM and NPERM models and also TRAN and NTRAN models. Specific *dL*-scores for permanent (dL_p) and transient (dL_t) predictions are calculated using equations 4 and 5, respectively.

$$\mathrm{dL_p}\left(\mathrm{i}\right)=\mathrm{L_{NPERM}}\left(\mathrm{i}\right)-\mathrm{L_{PERM}}\left(\mathrm{i}\right) \quad (4)$$

$$\mathrm{dL_t}\left(\mathrm{i}\right)=\mathrm{L_{NTRAN}}\left(\mathrm{i}\right)-\mathrm{L_{TRAN}}\left(\mathrm{i}\right) \quad (5)$$

The second step of BindML+ is to discriminate a predicted PBI site into either the permanent or the transient type using a Logistic Regression Model (LRM) (Bottom half of Figure 1). LRM performs binary classification by fitting a set of features using a logit function[23, 37, 38]. Features used in the LRM are based on difference between L_{PERM}(i) and L_{TRAN}(i) score, which is named the interface *type likelihood* (*tL*) score, computed for each residue in a predicted PBI site:

$$\mathrm{tL}\left(\mathrm{i}\right)=\mathrm{L_{PERM}}\left(\mathrm{i}\right)-\mathrm{L_{TRAN}}\left(\mathrm{i}\right), \quad (6)$$

Further, *tL*-scores are recast into Z-scores. A residue with a *tL* Z-score above zero it is more likely to be permanent, while a lower value below zero suggest that it is more likely to be transient.

A predicted PBI site in the previous step would consist of several surface patches, each of which contains 25 residues on average. We calculate the average *tL Z*-score of each patch in the predicted PBI site to identify two distinct patches: (1) a patch with the lowest average *tL Z*-score (min-patch) and (2) a patch with the highest average *tL* Z-score (max-patch). For the min-patch and the max-patch of the predicted PBI site, we compute the following five features each, thus ten features in total:

1. Average *tL Z*-score of residues in the min-patch/max-patch
2. Average *tL Z*-score of residues scoring above or equal to zero in the min-patch/max-patch
3. Average *tL Z*-score of residues scoring below zero in the min-patch/max-patch
4. The number of residues with *tL Z*-scores above or equal to zero in the min-patch/max-patch
5. The number of residues with *tL Z*-scores below zero in the min-patch/max-patch

The first three features concern average values of the *tL* Z-score and the latter two consider the number of residues with a certain range of *tL* Z-score in a patch. We performed leave-one-out cross-validation to train a LRM and make a prediction to a query protein with a predicted PBI that is left out from the training set. Several combinations of input features were tested. A probability computed by the LRM that is greater than or equal to a threshold value will classify a protein with a PBI that would be involved in permanent interaction. Otherwise, the probability that is lower than the set constant threshold value will classify a query protein as one that would participate in a transient interaction.

Evaluating PBI Site Prediction

The prediction performance of PBI residues was evaluated mainly using the Area Under the Curve (*AUC*) of Receiver Operating Characteristic (ROC) [24, 26, 27, 39] that plots sensitivity and specificity across multiple thresholds. 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. Using true positives (*TP*), which are the true PBI residues predicted correctly, true negatives (*TN*), which are non-PBI residues correctly classified, false positives (*FP*), which are false predictions of PBI site residues, and false negatives (*FN*), which are residues at PBI sites that are not predicted, the sensitivity and the specificity are define as

$$Sensitivity=\frac{TP}{TP+FN} \quad (7)$$

$$Specificity=\frac{TN}{TN+FP} \quad (8)$$

Evaluating PBI Site Classification

In evaluating PBI site classification, we considered the AUC of a plot of the true permanent classification rate (PCR) relative to the true transient classification rate (TCR). This is analogous to an AUC, where we compute the PCR (Eqn. 9) instead of the sensitivity (Eqn. 7) and the TCR instead of specificity. Concretely, we used the following equations to calculate *PCR* and *TCR*:

$$PCR=\frac{T_{Perm}}{T_{Perm}+F_{Perm}} \quad (9)$$

$$TCR=\frac{T_{Tran}}{T_{Tran}+F_{Tran}} \quad (10)$$

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

Results

Amino Acid and Secondary Structure Composition of Permanent and Transient Interfaces

To begin with, we compared the composition of amino acid frequencies of multiple sequence alignments (MSAs) of PBI sites and other surface regions (Fig. 2). Statistical significance of differences of amino acid fraction was tested by two-sample proportion test. Using the p-value cutoff of 0.05, the difference of fraction of almost all the amino acids were considered to be statistically significant, except for five amino acids, cysteine, leucine, glutamine, arginine, and valine (C, L, Q, R, V), at protein binding interfaces (Fig. 2A) and two amino acids, isoleucine and lysine (I, K), at non-protein binding interfaces (Fig. 2B). There are several notable differences as reported in a previous study[18, 23]. There is a clear bias in glycine and proline (G, P) composition at permanent interfaces compared to transient interfaces (Fig. 2A). Aromatic residues, tyrosine, tryptophan and phenylalanine residues (Y, W, P) are more abundant at permanent interfaces (Fig. 2A). The difference of amino acid proportion between permanent and transient proteins tends to be larger in binding interface (Fig. 2A) than in non-binding interface (Fig. 2B).

We further analyzed the secondary structure composition of permanent, transient binding interfaces as well as non-protein binding surfaces (Fig. 3). Protein surface consists of more loops (others), which is consistent with a work by Ansaris & Helms[23, 30]. We performed statistical tests to examine the statistical significance of the differences in fractions of secondary structures for the permanent, transient and non-interface observed in Figure 3. For the three classes of secondary structure, α-helices, β-strands, and others (loop), we performed the two-sample proportion test between pairs of permanent, transient, and non-interface surfaces. Statistical significance was observed secondary structure content of permanent interfaces. Using the p-value cutoff of 0.05, α-helix fraction in permanent interfaces is significantly lower than transient interfaces. For β-strands, difference of permanent interfaces and non-interface was significant. Permanent interfaces contain significantly more coils than transient interfaces and non-interfaces.

Analysis of Substitution Models

Next, we compared the constructed amino acid substitution matrices (Table II) using the Spearman rank correlation and the Kolmogorov-Smirnov (KS) distribution test [31, 40]. We found that PERM and TRAN have low correlation (*r*: 0.774) with each other, and is significantly different in their distribution (*D*: 0.105, *p*-value: < 0.05). On the other hand, as expected, NPERM and NTRAN matrices are not significantly different (*D*: 0.075, *p*-value: 0.211), which shows that mutations of residues on other surfaces (NPBI sites) in permanent and transient complexes are very similar. Further, the PERM with NPERM are significantly different shown by the KS test (*D*: 0.110, *p*-value: < 0.05).

Prediction of Permanent and Transient Interfaces

Using the PERM/NPERM and TRAN/NTRAN substitution models, we predicted PBI sites on the permanent and transient complexes in the JNT dataset (Fig. 4). The procedure employed was the first half of the entire procedure of BindML+ in Figure 1. A two-fold cross validation was carried out, i.e. for either permanent or transient protein sets in the combined JNT and Affinities dataset (mentioned in the Methods section), half of them were used to compute the substitution models (all protein sequences in the training set was used), which were then applied to the remaining half of the protein set to predict PBI sites. This procedure was repeated two times so that the two subsets of the dataset are handled as the testing set. Figure 4A shows the ROC curves on the permanent binding proteins, while prediction results on the transient proteins were shown in Figure 4B. For each case, predictions are performed using PERM/NPERM and TRAN/NTRAN substitution models.

The overall performance of the prediction was better on the permanent protein dataset (Fig. 4A) in comparison to the transient protein dataset (Fig. 4B). This result indicates that transient interfaces are more challenging to predict than permanent binding interfaces. The performance of PERM/NPERM substitution models and TRAN/NTRAN are similar, but the former showed slightly higher AUC than the latter for PBI site prediction for both permanent (Fig. 4A) and transient complexes (Fig. 4B). Interestingly, the permanent model performs better than transient models for the transient dataset (Fig. 4B) as well, by three percent AUC. Given that the permanent models performed better in both cases, in the following sections, we use the permanent substitution models to predict protein-protein interfaces in the first step of BindML+.

Examples of prediction of permanent and transient binding residues

In this section, we show examples of PBI site prediction and permanent- and transient-binding residues by BindML+. A permanent complex, cytoplasmic malate dehydrogenase (PDB code: 4MDH-A), is used as the first example. Figure 5A and 5B show the PBI site prediction for this protein. To detect the PBI site, PERM/NPERM substitution models were used in PHYML to calculate the *dL* score (Eqns. 3, 4). Residues with a negative *dL* score are predicted to be in the PBI site of the protein (Fig. 5A) and the predicted PBI site residues are mapped in black on its tertiary structure in Figure 5B.

The binding residue prediction is fairly successful for this protein with an AUC value of 0.828. In Figures 5C and 5D, the predicted binding residues are further classified into permanent or transient binding residues by applying the *tL Z*-score (Eqn. 6). Values above or equal to zero are permanent binding residue predictions (blue in Fig. 5D), whereas lower scores are predicted as transient binding residue predictions (red). Out of 103 predicted binding residues, 84 were classified as permanent binding residues, where as only 18 are as transient type. Thus, the majority of the residues are classified correctly as permanent; particularly, majority of actual binding residues are classified as permanent.

Four additional examples are shown in Figure 6. First two are examples of permanent complexes, while the latter two are transient complexes. The first protein is ascaris pepsin inhibitor-3 bound to porcine pepsin (PDB: 1F34). The PBI site of the pepsin inhibitor was reasonably well predicted with an AUC of 0.788. 56 predicted binding residues are dominated by permanent predictions (50 residues shown in blue) in agreement with the permanent nature of the interaction. Next example (Fig. 6B) is another permanent interface from staphostatin-staphopain complex (PDB: 1PXV). The AUC of the prediction for cysteine protease is 0.836. Among 73 residues predicted as binding, 63 were classified as permanent, whereas 10 are classified as transient residues. Actual binding residues are dominated as permanent prediction, and the misclassified 10 residues locate mainly far from the actual PBI sites.

The third example is a transient complex, a solution structure of cdc42 in complex with the GTPase binding domain of Wiskott-Aldrich Syndrome protein (PDB: 1CEE), whose binding interface is predicted with a specificity of 0.625 and an AUC of 0.607. In this transient example, we see much more transient type predictions at PBI sites as compared with the previous two examples (nine transient and eighteen permanent residue predictions). There are eight out of nine transient type residues in contact with the interacting partner. The last one is another transient complex, bovine β-lactoglobulin (PDB: 1BEB). 30 and 27 residues are predicted as transient and permanent, respectively. The overall PBI site prediction of this example has an AUC of 0.734. It is observed that predicted transient residues are clustered around the true interface, despite an appreciable mixture of permanent residue predictions at the periphery of the correct binding interface.

As we see in the examples in all the cases, both permanent and transient binding residues are predicted for permanent binding and transient binding proteins. However, permanent and transient binding proteins have difference in distributions of predicted residues. For permanent docking interfaces, a dominant number of residues are predicted as permanent (e.g. 81.8 to 86.3 % in Figs. 6A and 6B) and actual interfaces are occupied by the permanent type predictions. In the case of transient interfaces, residues that are predicted to participate in transient interaction do not share a dominating fraction in a true PBI, but are still clustered at the actual binding interface. The last step of BindML+ attempts to capture these differences and makes the final classification of a query protein to have either permanent or transient type interactions.

Classification of Protein-Protein Interface Predictions

Next, we discuss the results of the classification of permanent and transient protein interfaces, which is the second step of the BindML+ algorithm (Fig. 1). The final classification is made by employing LRM using the features computed for the max-patch and the min-patch (the patch which has the largest/smallest *tL Z*-score) in the predicted PBI site in a query protein. The performance shown in Figure 7 and Table II was computed by a leave-one-out cross-validation.

We compared the prediction performance of three sets of features for LRM: Using the average *tL Z*-scores of the max- and min-patches (i.e. the first three features listed in Methods section); the residue number count features of the max- and min-patches (i.e. the last two features in the list); and all the features. When the classification was applied to predicted PBI sites (Fig. 7A), using only the residue number features (the dotted line in the plot) showed a highly accurate prediction with an AUC of 0.957. It achieved a permanent classification rate (PCR) (Eqn. 9) of 0.941 and the transient classification rate (TCR) (Eqn. 10) of 0.907 when a *tL* score cutoff of 0.593 was used. Using all features (the solid line), a lower performance with an AUC of 0.793 was observed. The prediction with the average *tL Z*-score features showed lowest AUC value (0.725) among the three combinations.

We further tested the classification when the true protein interface is known (to assume that the interface residues are experimentally identified or perfectly predicted). Remarkably, all three combination features performed near perfect classification for permanent and transient interfaces with AUC at least greater than 0.950 (Table IV). In contrast to classification using the predicted interfaces (Table III), the *tL* Z-score based features performed with the near perfect classification (AUC: 0.991), while use of the residue counts feature performed slightly worse but still showed a significantly high AUC value (0.951). The near perfect classification for the cases when the true interface residues are known *a priori* vividly demonstrate that our strategy of using interaction type-specific substitution models is effective for differentiating permanent and transient interfaces.

Figure 8 illustrates the typical situation involving the min- and the max-patch as well as the other residues on the protein surface. The top panels (Figs. 8A and 8B) are results for tryptophan synthase complex, a permanent interaction complex. All patch *tL Z*-scores are mapped on to each representative surface residue shown is shown Fig 8A, where *tL Z*-score above zero (potential permanent residues) are colored cyan and *tL Z*-score values below zero (potential transient residues) are colored pink. Fig 8B shows that a subset of these residues that form the min-patch, which includes only four residues with transient predictions (red), whereas the max-patch contains 14 residues predicted as permanent. Eight residues in the max-patch (blue) are in direct contact with the interacting partner, whereas the min-patch has no contacting residues near the true interface. The bottom panels are results for a transient interacting structure, the αL I domain in complex with ICAM-1 (Figs. 8C and 8D). In this case, the min-patch (red) is closer to the true interface than the max-patch. There are three residues predicted to be transient in the min-patch that are in direct contact with the interacting partner.

The other patches between the maximum and the minimum *tL* Z-score value cover the remaining regions of predicted protein interface. In the example of the permanent interacting protein (Fig. 8B), 17 residues had positive *tL Z*-scores (cyan), while only five residues show negative *tL Z*-scores (pink). For permanent interacting proteins typically there are a dominant number of residues with a positive *tL Z*-score. On the other hand in the transient case (Fig. 8D), the true interface contains both permanent predictions (cyan) and clustered transient predictions in the min-patch (red). Thus, the residues that form the most negative scoring patch are a part of the true interface, demonstrating that the use of min-patches for transient interface classification is advantageous, especially when mixed with nearby residues of weaker permanent scores.

Discussion

We have developed a new computational algorithm, BindML+, which differentiates permanent and transient interaction types of predicted PBI sites on protein surfaces. Unlike several existing works that predict permanent or transient interaction types given a protein complex structure, BindML+ predicts permanent or transient interface in a single protein without knowledge of its interacting partner. Thus, BindML+ is the first method of its kind to be developed so far. Through cross-validation benchmarks on a large dataset of permanent and transient interacting protein complexes, our method was shown to perform very well in classifying between the two interaction types. When applied to known PBI sites, BindML+ classified them into two types with near perfect accuracy.

It is worthwhile to note that there are differences between permanent and transient PBIs when considering the distribution of residues predicted to be permanent and ones predicted to be transient by the *tL*-score. Both permanent and transient interacting proteins have both types of residues, ones predicted to be permanent and those that are predicted to be transient. In the case of permanent interacting proteins, the difference is that residues predicted to be part of a permanent complex are dominant overall, while for transient interacting proteins, residues predictions form local clusters on surfaces (i.e. the min-patch). Thus, capturing these characteristic features of permanent/transient PBIs has lead to successful prediction by BindML+.

The approach we developed here may also aid in the design of novel permanent or transient protein-protein interactions, particularly, changing permanent docking interfaces to transient type or vice versa by introducing interface mutations. 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 computed for various types of ligand binding sites, such as those that bind metals or small chemical ligands or cofactors, such as ATP, NAD, or GTP can be used in our framework for their prediction and classification.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Kastritis Panagiotis and Alexandre Bonvin for their assistance with the protein-protein binding affinities database. This work was supported in part by the National Institute of General Medical Sciences of the National Institutes of Health (R01GM075004, R01GM097528). DK also acknowledges funding from the National Science Foundation (IIS0915801, EF0850009, IOS1127027) and National Research Foundation of Korea Grant (NRF-2011-220-C00004).

References

1. Kastritis PL, Moal IH, Hwang H, Weng Z, Bates PA, Bonvin AMJJ, Janin J. A structure-based benchmark for protein-protein binding affinity. Protein Science. 2011; 20(3):482–491. [PubMed: 21213247]
2. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature. 2000; 403(6770):623–627. [PubMed: 10688190]
3. Ito T. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proceedings of the National Academy of Sciences. 2001; 98(8):4569–4574.
4. Giot L. A Protein Interaction Map of Drosophila melanogaster. Science. 2003; 302(5651):1727–1736. [PubMed: 14605208]
5. Xu Q, Canutescu AA, Wang G, Wang G, Shapovalov M, Shapovalov M, Obradovic Z, Obradovic Z, Dunbrack RL Jr, Dunbrack RL Jr. Statistical Analysis of Interface Similarity in Crystals of Homologous Proteins. J Mol Biol. 2008; 381(2):487–507. [PubMed: 18599072]
6. Herberg FW, Dostmann WR, Zorn M, Davis SJ, Taylor SS. Crosstalk between domains in the regulatory subunit of cAMP-dependent protein kinase: influence of amino terminus on cAMP binding and holoenzyme formation. Biochemistry. 1994; 33(23):7485–7494. [PubMed: 8003514]
7. Gegner JA, Dahlquist FW. Signal transduction in bacteria: CheW forms a reversible complex with the protein kinase CheA. Proc Natl Acad Sci USA. 1991; 88(3):750–754. [PubMed: 1992467]
8. Lambright DG, Sondek J, Bohm A, Skiba NP, Hamm HE, Sigler PB. The 2.0 Å crystal structure of a heterotrimeric G protein. Nature. 1996; 379(6563):311–319. [PubMed: 8552184]
9. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR, Bateman A. The Pfam protein families database. Nucleic Acids Res. 2009; 38(Database):D211–D222. [PubMed: 19920124]
10. Krause R, Mering von C, Bork P, Dandekar T. Shared components of protein complexes--versatile building blocks or biochemical artefacts? Bioessays. 2004; 26(12):1333–1343. [PubMed: 15551274]
11. Han JDJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJM, Cusick ME, Roth FP, Vidal M. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature. 2004; 430(6995):88–93. [PubMed: 15190252]
12. Fierens K, Gils A, Sansen S, Brijs K, Courtin CM, Declerck PJ, De Ranter CJ, Gebruers K, Rabijns A, Robben J, Campenhout S, Volckaert G, Delcour JA. His374 of wheat endoxylanase inhibitor TAXI-I stabilizes complex formation with glycoside hydrolase family 11 endoxylanases. FEBS J. 2005; 272(22):5872–5882. [PubMed: 16279951]
13. Stratikos E, Gettins PG. Major proteinase movement upon stable serpin-proteinase complex formation. Proc Natl Acad Sci USA. 1997; 94(2):453–458. [PubMed: 9012804]
14. Olson MW, Gervasi DC, Mobashery S, Fridman R. Kinetic analysis of the binding of human matrix metalloproteinase-2 and -9 to tissue inhibitor of metalloproteinase (TIMP)-1 and TIMP-2. J Biol Chem. 1997; 272(47):29975–29983. [PubMed: 9368077]
15. Kiel C, Selzer T, Shaul Y, Schreiber G, Herrmann C. Electrostatically optimized Ras-binding Ral guanine dissociation stimulator mutants increase the rate of association by stabilizing the encounter complex. Proc Natl Acad Sci USA. 2004; 101(25):9223–9228. [PubMed: 15197281]
16. van der Merwe PA, Barclay AN, Mason DW, Davies EA, Morgan BP, Tone M, Krishnam AK, Ianelli C, Davis SJ. Human cell-adhesion molecule CD2 binds CD58 (LFA-3) with a very low

affinity and an extremely fast dissociation rate but does not bind CD48 or CD59. Biochemistry. 1994; 33(33):10149–10160. [PubMed: 7520278]

17. Maenaka K, van der Merwe PA, Stuart DI, Jones EY, Sondermann P. The human low affinity Fcgamma receptors IIa, IIb, and III bind IgG with fast kinetics and distinct thermodynamic properties. J Biol Chem. 2001; 276(48):44898–44904. [PubMed: 11544262]
18. La D, Kihara D. A novel method for protein-protein interaction site prediction using phylogenetic substitution models. Proteins. 2011
19. Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C. Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. Structure. 2010; 18(10):1233–1243. [PubMed: 20947012]
20. Nooren IMA. NEW EMBO MEMBER'S REVIEW: Diversity of protein-protein interactions. EMBO J. 2003; 22(14):3486–3492. [PubMed: 12853464]
21. Ezkurdia I, Bartoli L, Fariselli P, Casadio R, Valencia A, Tress ML. Progress and challenges in predicting protein-protein interaction sites. Briefings in Bioinformatics. 2008; 10(3):233–246. [PubMed: 19346321]
22. Mintseris J. Structure, function, and evolution of transient and obligate protein-protein interactions. Proceedings of the National Academy of Sciences. 2005; 102(31):10930–10935.
23. Ansari S, Helms V. Statistical analysis of predominantly transient protein-protein interfaces. Proteins. 2005; 61(2):344–355. [PubMed: 16104020]
24. Block P, Block P, Paern J, Paern J, Hüllermeier E, Hüllermeier E, Sanschagrin P, Sanschagrin P, Sotriffer CA, Sotriffer CA, Klebe G, Klebe G. Physicochemical descriptors to discriminate protein-protein interactions in permanent and transient complexes selected by means of machine learning algorithms. Proteins. 2006; 65(3):607–622. [PubMed: 16955490]
25. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci. 2004; 13(1):190–202. [PubMed: 14691234]
26. Mintseris J, Weng Z. Atomic contact vectors in protein-protein recognition. Proteins. 2003; 53(3): 629–639. [PubMed: 14579354]
27. Liu R, Jiang W, Zhou Y. Identifying protein–protein interaction sites in transient complexes with temperature factor, sequence profile and accessible surface area. Amino Acids. 2009; 38(1):263–270. [PubMed: 19214704]
28. Neuvirth H, Raz R, Schreiber G. ProMate: A Structure Based Prediction Program to Identify the Location of Protein–Protein Binding Sites☆. J Mol Biol. 2004; 338(1):181–199. [PubMed: 15050833]
29. Chen H, Zhou HX. Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against NMR data. Proteins. 2005; 61(1):21–35. [PubMed: 16080151]
30. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci USA. 1996; 93(1):13–20. [PubMed: 8552589]
31. Nooren IMA, Thornton JM. Structural characterisation and functional significance of transient protein-protein interactions. J Mol Biol. 2003; 325(5):991–1018. [PubMed: 12527304]
32. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000; 28(1):235–242. [PubMed: 10592235]
33. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32(5):1792–1797. [PubMed: 15034147]
34. Miller S, Janin J, Lesk AM, Chothia C. Interior and surface of monomeric proteins. J Mol Biol. 1987; 196(3):641–656. [PubMed: 3681970]
35. Jones D, Taylor W, Thornton J. The rapid generation of mutation data matrices from protein sequences. Bioinformatics. 1992
36. Henikoff S, Henikoff J. Amino Acid Substitution Matrices from Protein Blocks. Proceedings of the National Academy of Sciences. 1992
37. Hosmer, DW.; Lemeshow, S. Applied logistic regression. Wiley-Interscience; 2000.
38. Hilbe, J. Logistic regression models. Chapman & Hall/CRC; 2009.

39. Gribskov M, Robinson NL. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Computers & chemistry. 1996; 20(1):25–33. [PubMed: 16718863]

40. Tseng YY. Estimation of Amino Acid Residue Substitution Rates at Local Spatial Regions and Application in Protein Function Inference: A Bayesian Monte Carlo Approach. Mol Biol Evol. 2005; 23(2):421–436. [PubMed: 16251508]

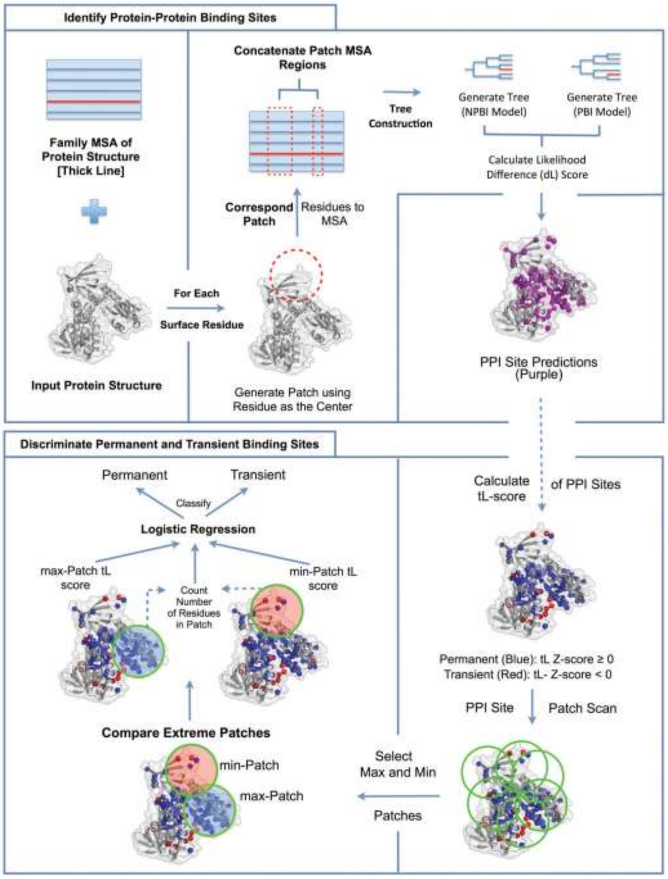
Identify Protein-Protein Binding Sites
Family MSA of Protein Structure [Thick Line]
Input Protein Structure
For Each
Surface Residue
Generate Patch using Residue as the Center
Correspond Patch
Residues to MSA
Concatenate Patch MSA Regions
Tree Construction
Generate Tree (NPBI Model)
Generate Tree (PBI Model)
Calculate Likelihood Difference (dL) Score
PPI Site Predictions (Purple)
Discriminate Permanent and Transient Binding Sites
Calculate tL-score
of PPI Sites
Permanent (Blue): tL Z-score ≥ 0
Transient (Red): tL- Z-score < 0
PPI Site
Patch Scan
Select Max and Min
Patches
Compare Extreme Patches
min-Patch
max-Patch
max-Patch tL score
min-Patch tL score
Count Number of Residues in Patch
Logistic Regression
Classify
Permanent
Transient

Figure 1.

Flowchart of the BindML+ method for classifying protein interface predictions.

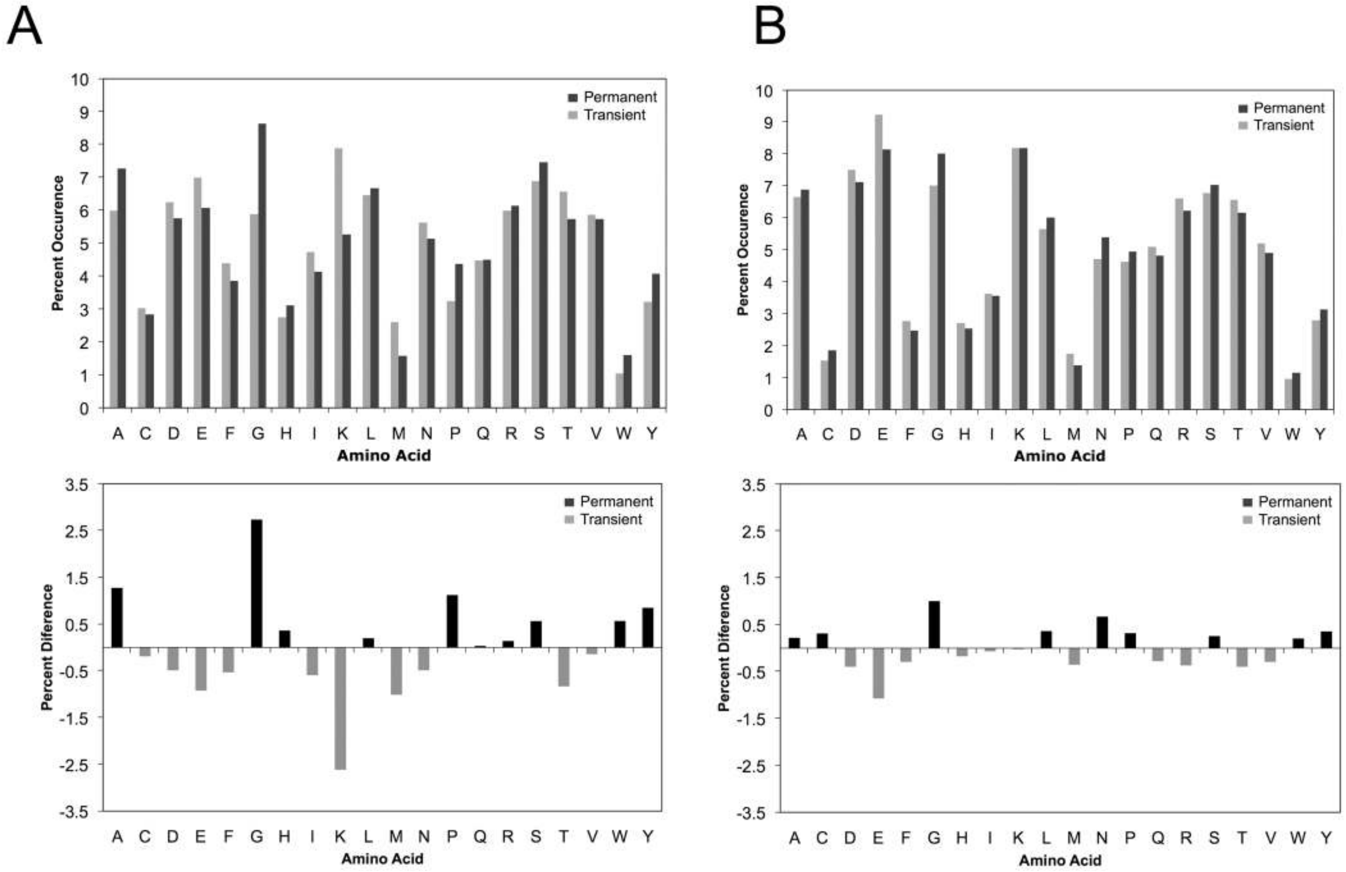

Figure 2.

Amino acid frequencies of **A**, interface and **B**, non-interface regions in permanent and transient complexes.

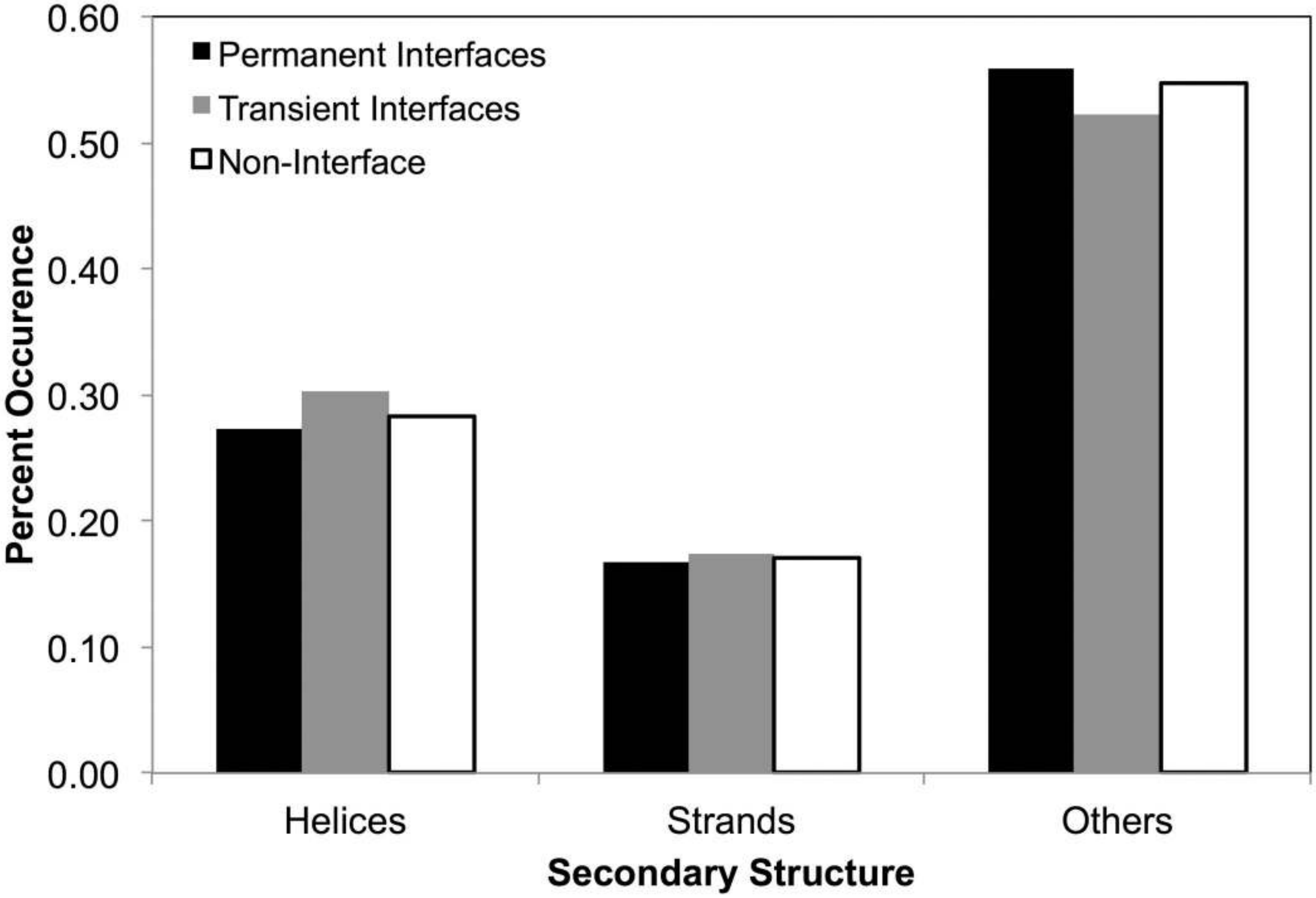

Figure 3.

Comparison of secondary structure composition of permanent and transient interfaces. The statistical significance of the two-sample test for proportions of amino acids and the standard errors were computed assuming the secondary structure fraction follows the normal distribution.

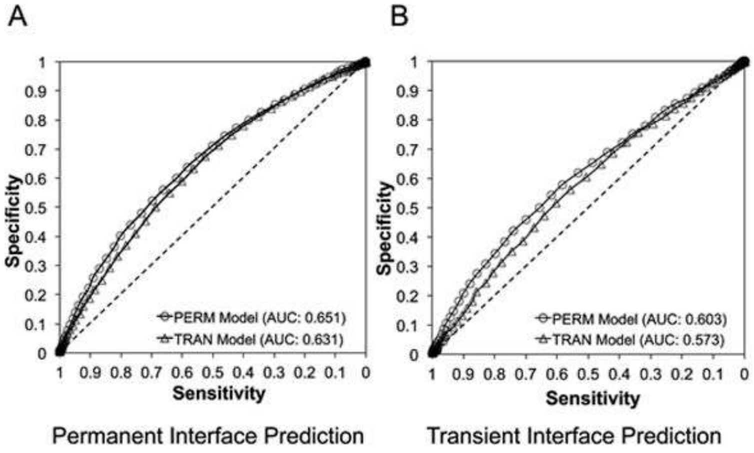

Figure 4.

The ROC curve for the interface prediction benchmark results on the combination of the JNT and Affinities dataset. **A**, Permanent PBI site prediction performance is shown using the PERM/NPERM model in open circles; **B**, transient PBI site prediction performance using the TRAN/NTRAN model is shown in open triangles. The dashed line indicates expected performance of random predictions.

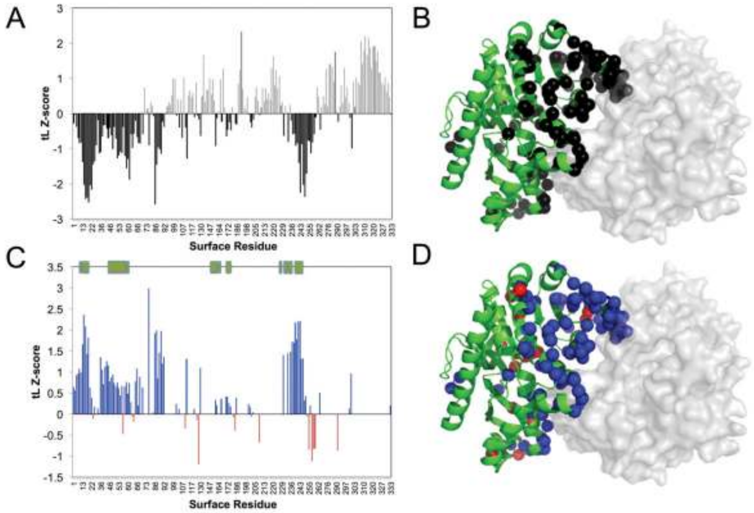

Figure 5.

Examples of binding site scores mapped to surface residues. PBI site predictions and classification is performed for cytoplasmic malate dehydrogenase (PDB: 4MDH-A), the structure in green, while the interacting partner in translucent grey surface. **A**, Distribution of *dL* Z-scores, where residues predicted as interface are shown as in thick black bars; **B**, predicted interface residues are shown in black spheres; **C**, Distribution of *tL Z*-scores of the PBI site predictions, where blue are permanent site predictions while red are transient site predictions; **D**, *tL Z*-scores mapped to the PDB structure by their corresponding colors.

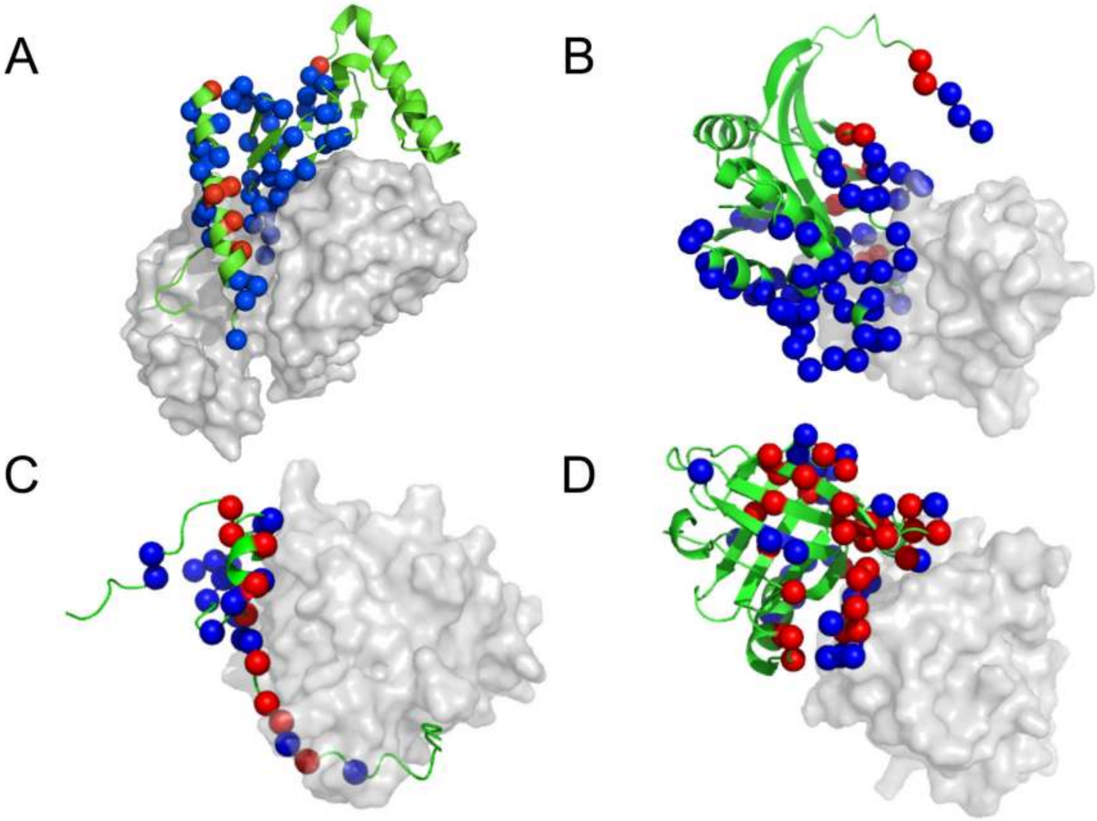

Figure 6.

Examples of *tL Z*-scores mapped to structures. PBI site predictions and classification is performed on the structure in green, while the interacting partner in translucent grey surface. Residues with a positive *tL Z*-score (i.e. residues predicted to be permanent interaction) is shown in blue and those with a negative *tL Z*-score (i.e. transient interaction) are shown in red. **A**, Permanent interaction (PDB: 1F34-B): structure of ascaris pepsin inhibitor-3 bound to porcine pepsin; **B**, Permanent interaction (PDB: 1PXV-A) staphostatin-staphopain complex, a forward binding inhibitor in complex with its target cysteine protease; **C**, Transient interaction (PDB: 1CEE-B): solution structure of cdc42 in complex with the GTPase binding domain of wasp; **D**, Transient interaction (PDB: 1BEB-A): bovine beta-lactoglobulin.

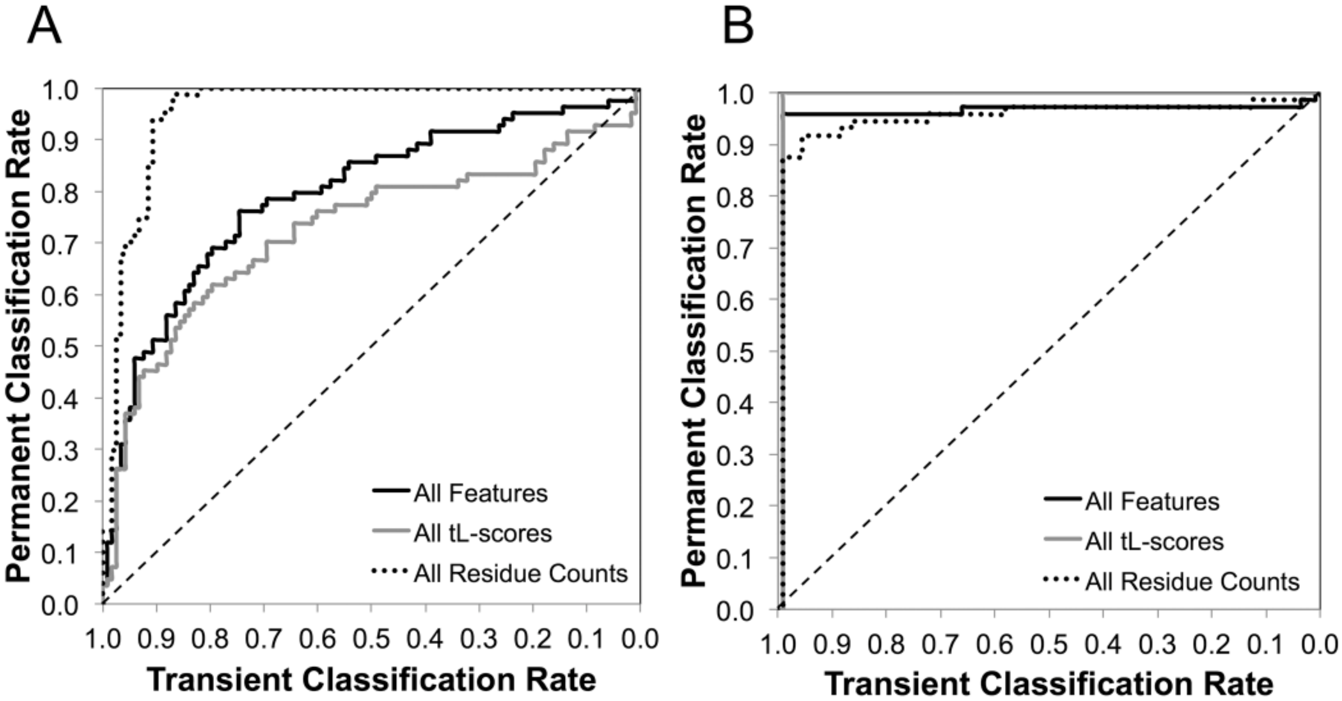

Figure 7.

ROC curves of the permanent and transient protein interface classification benchmark for **A**, predicted and **B**, known interfaces. ROC curves using the all features are shown in solid black curve, prediction results using average *tL Z*-scores are shown in gray (1-3 in the feature list), and prediction results using the residue counts (4, 5 in the feature list) are shown in dotted line.

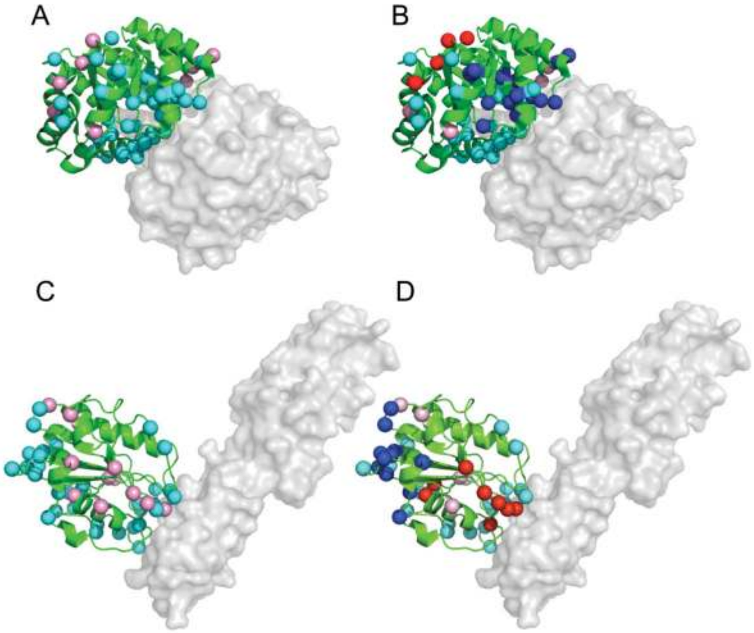

Figure 8.

A, Examples of surface residues scores and **B**, their corresponding in min- and max-patches for: Tryptophan synthase complex from a hyperthermophile (PDB: 1WDW-C); **C**, Colored surface residues of the predicted interface and **D**, min- and max-patches for: AlphaL I domain in complex with ICAM-1 (PDB: 1MQ8-B). Residues from patches with *more positive scores* are colored in cyan, whereas those with *more negative scores* are colored in pink. The residues in the min-patch are colored in red, while those in the max-patch are colored in blue for each structural example.

Table I

A list of number of complexes with K_d values in the Affinities Dataset.

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

Table II

Substitution matrices of protein binding and non-binding interfaces.

A. Log odds amino acid substitution matrix for permanent protein binding interface (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	-1	0	0	-2	-2	0
R	-1	3	-1	-1	-2	0	-1	-1	-1	-1	-1	1	-1	-2	-1	-1	-1	-2	-2	-1
N	-1	-1	3	0	-2	-1	-1	-1	0	-2	-1	0	-1	-2	-1	0	-1	-1	-1	-1
D	-1	-1	0	3	-3	-1	1	-1	-1	-2	-2	-1	-1	-2	-1	-1	-1	-2	-2	-2
C	-1	-2	-2	-3	5	-2	-2	-2	-2	-2	-2	-2	-2	-2	-3	-1	-2	-2	-2	-1
Q	-1	0	-1	-1	-2	3	0	-1	0	-1	-1	0	-1	-2	-1	-1	-1	-2	-2	-1
E	0	-1	-1	1	-2	0	3	-1	-1	-2	-1	0	-1	-2	-1	-1	-1	-3	-2	-1
G	-1	-1	-1	-1	-2	-1	-1	3	-1	-2	-2	-1	-2	-2	-2	-1	-2	-2	-2	-2
H	-1	-1	0	-1	-2	0	-1	-1	4	-2	-1	-1	-1	-1	-1	-1	-1	-1	0	-2
I	-1	-1	-2	-2	-2	-1	-2	-2	-2	3	1	-1	0	0	-1	-1	-1	-2	-1	1
L	-1	-1	-1	-2	-2	-1	-1	-2	-1	1	3	-1	1	0	-1	-1	-1	-1	-1	0
K	-1	1	0	-1	-2	0	0	-1	-1	-1	-1	3	-1	-2	-1	-1	-1	-2	-2	-1
M	-1	-1	-1	-1	-2	-1	-1	-2	-1	0	1	-1	4	0	-1	-1	0	-1	-1	0
F	-1	-2	-2	-2	-2	-2	-2	-2	-1	0	0	-2	0	4	-2	-1	-1	0	1	0
P	-1	-1	-1	-1	-3	-1	-1	-2	-1	-1	-1	-1	-1	-2	3	-1	-1	-2	-2	-1
S	0	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	2	0	-2	-1	-1
T	0	-1	-1	-1	-2	-1	-1	-2	-1	-1	-1	-1	0	-1	-1	0	3	-2	-2	0
W	-2	-2	-1	-2	-2	-2	-3	-2	-1	-2	-1	-2	-1	0	-2	-2	-2	5	0	-2
Y	-2	-2	-1	-2	-2	-2	-2	-2	0	-1	-1	-2	-1	1	-2	-1	-2	0	4	-1
V	0	-1	-1	-2	-1	-1	-1	-2	-2	1	0	-1	0	0	-1	-1	0	-2	-1	3

B. Amino acid substitution matrix for permanent non-protein binding interface (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

B. Amino acid substitution matrix for permanent non-protein binding interface (NPERM).

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
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

C. Amino acid substitution matrix for transient protein binding interface (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

C. Amino acid substitution matrix for transient protein binding interface (TRAN).

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
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

D. Amino acid substitution matrix for transient non-protein binding interface (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

Table III

Classification of predicted protein-protein interfaces.

Combination	PCR	TCR	AUC
All *tL Z*-scores	0.702	0.695	0.725
All residue counts	0.941	0.907	0.957
All Features	0.762	0.746	0.793

PCR and TCR values are reported for the *tL* score threshold that gives the closest point to the perfect prediction of TCR=1.0 and PCR = 1.0.

Table IV

Classification of true protein-protein interfaces.

Combination	PCR	TCR	AUC
All *tL Z*-scores	1.000	0.991	0.991
All residue counts	0.917	0.955	0.951
All Features	0.958	0.991	0.960