Prediction of Interface Residues in Protein–Protein Complexes by a Consensus Neural Network Method: Test Against NMR Data

Huiling Chen[1] **and Huan-Xiang Zhou**[2]

[1]*Department of Physics, Drexel University, Philadelphia, Pennsylvania*

[2]*Department of Physics and Institute of Molecular Biophysics and School of Computational Science, Florida State University, Tallahassee, Florida*

ABSTRACT **The number of structures of protein–protein complexes deposited to the Protein Data Bank is growing rapidly. These structures embed important information for predicting structures of new protein complexes. This motivated us to develop the PPISP method for predicting interface residues in protein–protein complexes. In PPISP, sequence profiles and solvent accessibility of spatially neighboring surface residues were used as input to a neural network. The network was trained on native interface residues collected from the Protein Data Bank. The prediction accuracy at the time was 70% with 47% coverage of native interface residues. Now we have extensively improved PPISP. The training set now consisted of 1156 nonhomologous protein chains. Test on a set of 100 nonhomologous protein chains showed that the prediction accuracy is now increased to 80% with 51% coverage. To solve the problem of over-prediction and under-prediction associated with individual neural network models, we developed a consensus method that combines predictions from multiple models with different levels of accuracy and coverage. Applied on a benchmark set of 68 proteins for protein–protein docking, the consensus approach outperformed the best individual models by 3–8 percentage points in accuracy. To demonstrate the predictive power of cons-PPISP, eight complex-forming proteins with interfaces characterized by NMR were tested. These proteins are nonhomologous to the training set and have a total of 144 interface residues identified by chemical shift perturbation. cons-PPISP predicted 174 interface residues with 69% accuracy and 47% coverage and promises to complement experimental techniques in characterizing protein–protein interfaces. Proteins 2005;61:21–35.**

© 2005 Wiley-Liss, Inc.

Key words: protein–protein interaction; protein complexes; neural network; interface prediction; protein docking

INTRODUCTION

Many biological processes are carried out, or regulated, through the interactions between proteins. Genome-wide two-hybrid analysis[1,2] shows that the vast majority of proteins have interacting partners in the cell, and often more than one. Therefore residues in protein–protein interfaces are essential for protein function. Many important applications follow directly from the identification of interface residues, such as drug design, protein mimetics engineering, elucidation of molecular pathways,[3,4] and understanding of disease mechanisms.[5] The proper identification of interface residues can also guide the docking process to build the structural model of protein–protein complexes.[6] These considerations motivated us to develop a method, Protein–Protein Interaction Site Predictor (PPISP), for predicting interface residues based on unbound protein structures and characteristics of interface residues.[7] We have now made significant improvements on PPISP, to the extent that the method now promises to complement NMR and other experimental techniques in characterizing protein-protein interfaces. Here we report this improved method, cons-PPISP.

The structures of protein–protein complexes in the Protein Data Bank (PDB) embed important information for predicting interaction sites of protein–protein complexes. Many studies have investigated the characteristics of interfaces in different types of complexes, such as homodimers versus heterodimers, enzyme-inhibitor complexes, antigen–antibody complexes, transient complexes versus obligatory complexes, large interfaces versus small interfaces.[7–23] Although conclusions were sometimes conflicting, some common features can be extracted. For example, evolutionarily important residues tend to be spatial clustered,[15,17–19,22,23] and nonpolar residues are favored whereas charged and polar residues (except for Arg) are disfavored in protein interfaces.[7] These characteristics formed the basis of the PPISP method.

In PPISP, the position-specific sequence profiles produced by PSI-BLAST[24] and solvent accessibility of spa-

Grant sponsor: the National Institutes of Health; Grant number: GM58187.

*Correspondence to: Huan-Xiang Zhou, Department of Physics and Institute of Molecular Biophysics and School of Computational Science, Florida State University, Tallahassee, FL 32306; E-mail zhou@sb.fsu.edu

Received 4 January 2005; Revised 22 February 2005; Accepted 24 February 2005

Published online 3 August 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20514

tially neighboring surface residues were used as input to a neural network. The neural network approach has been shown to be quite successful in predicting protein secondary structure and solvent accessibility.[25–27] Sequence profiles of residue substitutions in naturally evolved protein families are highly specific for details of a particular protein structure. The use of sequence profiles has been shown to be a key in improving secondary structure prediction.[25] PSI-BLAST sequence profiles are exhaustive, convenient to use, and can be easily expanded to large-scale applications. The neural network of PPISP was trained on native interface residues. The training set was collected from 678 nonhomologous (sequence identity $< 40\%$) complex-forming protein chains in the PDB. The prediction accuracy at the time was 70%, and the predictions covered 47% of native interface residues. A number of similar methods have since been published.[28–33]

With the rapid increase of structures of protein–protein complexes deposited in the PDB, we wondered whether the PPISP method can be improved by training on a more exhaustive and less redundant data set. The training set has now been expanded to 1156 nonhomologous protein chains with $< 30\%$ sequence identity. When this set was trained using the same protocol as before, prediction accuracy indeed increased significantly, to 86%. However, at the same time the coverage of native interface residues also downgraded significantly, to just 17%. It was apparent that a better compromise between prediction accuracy and coverage had to be found. Higher accuracy means that there is only a small number of reliable predictions are made, thus a large number of potential interface residues are missed and coverage of native interface residues is reduced. Neither accuracy nor coverage alone constitutes a good measure of performance. The results of CAPRI (Critical Assessment of PRedicted Interactions)[34] indicate that a good prediction requires at least half of native interface residues to be correctly identified. Therefore our goal was to increase the coverage to 50% and achieve as high an accuracy as possible. After a number of refinements of the PPISP method, we obtained a prediction accuracy of 80% at 51% coverage.

Close examination of the predictions for individual proteins in the test set revealed another problem. For some proteins interface residues were over-predicted, but for others interface residues were under-predicted or not predicted at all. For a given protein, neither over-prediction nor under-prediction is desirable, even if the collective accuracy and coverage measures on a set of proteins are good. This problem persisted when we varied the neural network models. To solve this problem, we constructed a series of models ranging from high accuracy/low coverage to low accuracy/high coverage. We developed a consensus approach based on these models. Predicted interface residues were ranked by consensus score and clustered according to their locations on the protein surface. If there was a large number of predictions with high consensus scores, collection of interface residues was limited to these (to prevent over-prediction), otherwise the process was extended to predictions with lower consensus scores (to avoid under-prediction). Test on a benchmark set of 68 proteins for protein–protein docking[35] showed that the consensus approach outperformed the best individual models by 3–8 percentage points in accuracy. The predictive power of cons-PPISP was further demonstrated on eight complex-forming proteins with interfaces characterized by NMR. Of a total of 144 interface residues identified by chemical shift perturbation, cons-PPISP predicted 174 interface residues with 69% accuracy and 47% coverage. This much improved PPISP method promises to complement NMR and other experimental techniques in characterizing protein–protein interfaces.

MATERIALS AND METHODS

Generation of the Data Set

The strategy for collecting interface data to train and test neural networks was essentially the same as in our previous work.[7] We compiled an exhaustive nonhomologous set of protein complexes by examining all multiple-chain protein entries in the PDB (Jan 2002 release). Excluded were chains shorter than 40 residues as well as pairs of chains with less than 20 residues in interfacial contact on either side. Interfacial contact was defined as a pair of heavy atoms from two sides of an interface that are within 5 Å. For PDB entries with more than two protein chains, each chain was assigned at most one partner, and the interface chosen was the one with the most residues forming interfacial contacts. A pair of protein chains sharing an interface is called a dimer throughout this paper. Each of the collected sequences was aligned against all the other sequences by PSI-BLAST.[24] All chains with high homologies were collected in a cluster and one representative chain was chosen from each cluster with the provision that heterodimers had higher priority than homodimers and longer chains had higher priority than shorter ones. The cutoff for high homology this time was set at 30% identity over the aligned region, which had to cover at least 90% of either of the two sequences. A homodimer was identified as a pair of protein chains that shared an interface and satisfied the criteria: (1) over 90% of both chains were aligned and (2) the sequence identity over the aligned region was at least 95%.

The resulting nonhomologous set consisted of 1256 chains. Within the data set, 798 were chains from homodimers and 458 were chains from heterodimers. From these a total of 225,237 surface residues were collected using the criterion of at least 10% surface exposure to solvent. Surface area was calculated using the DSSP program.[36] Of the surface residues, 52,623 (or 23%) were classified as interface residues, with at least one interfacial contact. This is called the "thrd = 1" criterion. In our previous study we used the "thrd = 3" criterion in training neural networks. We chose 100 chains as the test set, of which 58 chains were from heterodimers and 42 from homodimers. The remaining 1156 chains were used as the training set. These statistics are listed in Table I.

In our treatment, each protein formed only one interface. This was the one involving the most interfacial residues, denoted as the major interface. In PDB entries with more than two protein chains, a chain can potentially form more than one interface. Residues in these other interfaces were either classified as noninterface residues (leading to the interface statistics in Table I) or eliminated from the training set altogether. The latter treatment is denoted as "intfM." In our data set of 1256 protein chains, 727 had a single partner and thus actually formed dimers. The number of chains in contact with 2, 3, 4, 6, 7, or more other chains was 235, 182, 49, 43, 9, 7, and 4, respectively. Note that these calculations were done using the original PDB entries, thus crystal contacts were included and no specific consideration was given to the biologically significant oligomeric state. There were a total of 10,972 residues in "minor" interfaces.

Neural Network Architecture

Two feed-forward, back-propagation neural networks were used consecutively as before. In the previous study, the first network had 21 × 20 input nodes, in which the first quantity was the number (i.e., 20) of entries in a sequence profile plus one for solvent accessibility, and the second quantity was the window size, i.e., one for the residue under consideration plus 19 for its spatially nearest neighbors. In this study it was found decreasing the window size from 20 to 15 improved performance of the network. In the final version the first network had an input layer with 21 × 15 nodes, a hidden layer with 150 nodes, and an output layer with 2 nodes. The input layer of the second network had 3 × 15 nodes, in which the first quantity is two for the output values of the first network plus one for solvent accessibility. The second network was completed with 30 hidden nodes and two output nodes. The predictor was trained at different learning rate and a value of 0.001 that gave optimal performance was selected.

Assessment of Predictions

To assess the predictions, two quantities were calculated. Accuracy was defined as the percentage of correctly predicted interface residues among all predictions. A prediction was considered as correct if it was either a native interface residue or among the four nearest spatial neighbors of a native interface residue. If the neighbors were not counted, the percentage of correct predictions was referred to as "strict" accuracy. Coverage was defined as the fraction of native interface residues predicted among all native interface residues.

Consensus Approach Based on Different Neural-Network Models

The consensus approach consisted of two steps: (1) clustering of all predictions from different neural network models, and (2) selecting a cluster or clusters as the final predictions. In the first step, each predicted residue was assigned a consensus score, defined as the number of models making such a prediction. The predictions were then sorted according to consensus score. Starting with the batch of predictions with the highest consensus score, residues were clustered if they were among the four nearest neighbors of each other. At the end if all the cluster sizes were less than 20, then the next batch of predictions with the second highest consensus score was used to grow the clusters and add new clusters. The process was continued until one cluster size went beyond 20 or all the predictions have been clustered.

In the second step, clusters were selected according to the following considerations. First, clusters with a single residue were eliminated. Second, if the highest consensus score of a cluster was higher than that of another cluster by six or more, then the cluster with the low score was eliminated. If the highest consensus scores of two clusters differed by less than six, then they were not differentiated by consensus score, but were compared in size. Third, among all the retained clusters with comparable consensus scores, the largest cluster was automatically selected, and other clusters were also selected if their sizes were smaller by less than five.

RESULTS AND DISCUSSION

Characteristics of Interface Residues

In our previous study,[7] analysis of interface characteristics showed that overall nonpolar residues were favored in the interface over charged and polar residues, and interface residues were more conserved than noninterface surface residues. Here these characteristics were further analyzed, both to gain better understanding of protein interfaces and to help improve the prediction method. In addition to hydrophobicity and conservation, we also sought to detect subtle differences in solvent accessibility between interface and noninterface residues on the updated data set.

The analysis on the updated data set reinforced previous findings on interface hydrophobicity and conservation. Figure 1(a) clearly shows that nonpolar residues (Leu, Ile, Phe, Tyr, Val, and Met) had higher populations in the interface collection than in the noninterface collection, while charged residues, except for Arg, had lower populations in the interface collection. The diagonal element of a sequence profile produced by PSI-BLAST signifies the mutability of each residue along the sequence. The higher the element, the less frequent its mutation (i.e., the more conserved the residue). The average values of the diagonal elements in the interface and noninterface collections for each type of residues are displayed in Figure 1(b). For all the residue types, the conservation scores in the interface collection were higher than in the noninterface collection.

Work of Jones and Thornton[10] has suggested that interface residues are more solvent accessible than noninterface surface residues. The average solvent accessibility results, calculated over the interface and noninterface collections separately, demonstrate a significant difference in solvent accessibility between these two collections.

The results in Figure 1 justify the use of sequence profiles and solvent accessibility as input data in our neural network approach for predicting interface residues. Sequence profiles can capture the differential characteristics of interface and noninterface residues in hydrophobicity and conservation, and the inclusion of solvent accessibility as part of the input adds a putatively independent

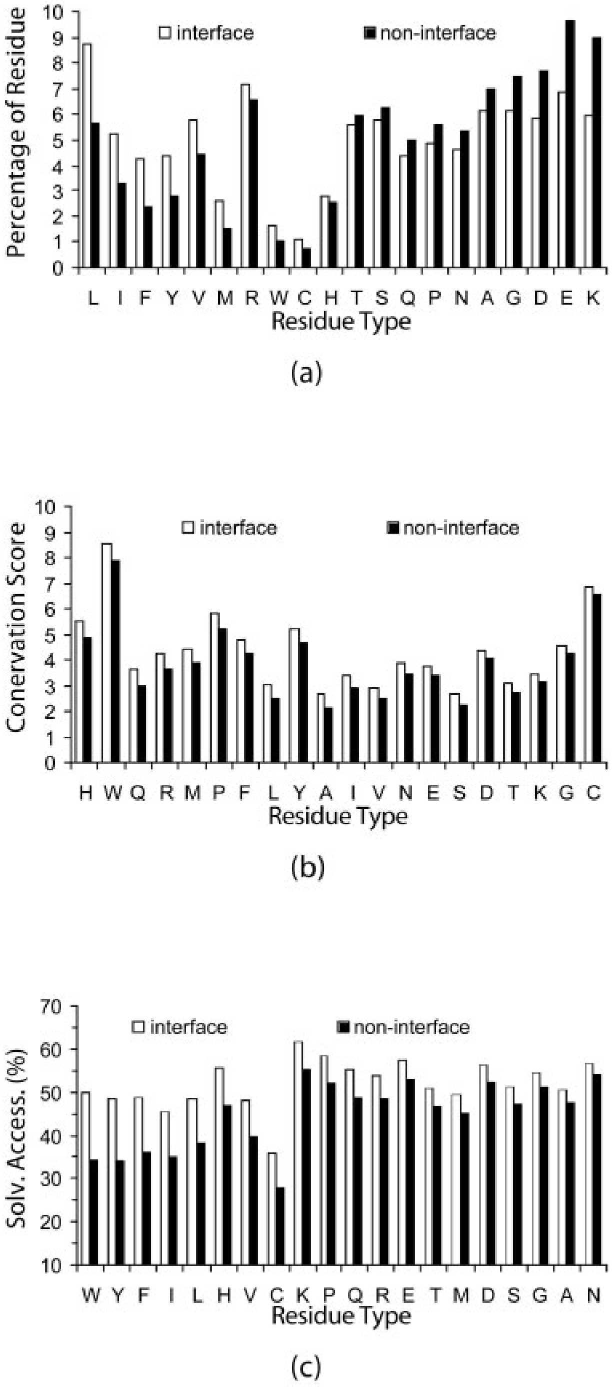

Fig. 1. Characteristics of interface residues compared to noninterface surface residues. (**a**) Percentages of the 20 types of amino acids in the interface and noninterface collections. The abscissa is in descending order of the difference between the two collections. (**b**) Average conservation scores in the interface and non-interface collections for the 20 types of amino acids, in descending order of the difference. (**c**) Average relative solvent accessibilities in the interface and noninterface collections for the 20 types of amino acids, in descending order of the difference. Results were obtained from analysis of surface residues in the training set.

measure of discrimination between the two types of surface residues. These differences between interface and noninterface residues are also easily rationalized. Since interface residues will be buried upon the binding of a partner protein, they are akin to interior residues, which tend to be more hydrophobic. The portion of the surface of an interface residue that becomes buried by the partner protein is counted as exposed. This counting increases the solvent accessibility and explains the difference with noninterface residues on the protein surface. Interfaces are where evolutionarily important residues are most likely found, and these residues are most likely to be conserved.

Comparison of the Updated and Previous Data Sets

The updated data set contained 1256 nonhomologous protein chains, of which 798 were from homodimers and 458 (or, 36.5%) were from heterodimers (Table I). Within the new training set of 1156 chains, 34.6% were from heterodimers. In comparison, the previous data set consisted of 924 chains, of which 360 (or, 39.0%) were from heterodimers, but just 18.6% of the 678 chains used for training were from heterodimers. With the maximal level of identity lowered from 40% to 30%, the new data set was also much less redundant, and covered much greater regions in sequence space. The increased coverage in sequence space allowed for further analysis of potential differences within the new data set. Figure 2(a) shows the distribution of the proteins according to their surface areas and interface areas. The data set was concentrated in small proteins and relatively sparse in large proteins, even though larger proteins were given higher priority in the collection. There was no apparent difference between homodimers and heterodimers in the distribution of surface and interface areas. The fraction, $f_{\rm intf}$, of interface residues among all surface residues in a protein tended to decrease with the protein size, as shown by Figure 2(b). When the number of surface residues, $N_{\rm surf}$, was sorted and averages in batches of 50 were calculated for $N_{\rm surf}$ and $f_{\rm intf}$, the following relation was obtained:

$$\bar{f}_{\rm intf} = 6.7\bar{N}_{\rm surf}{}^{-0.65} \qquad (1)$$

The decrease of $f_{\rm intf}$ with increasing $N_{\rm surf}$ motivated us to consider dividing the data set into three subsets: small proteins, those with surface areas less than 10,000 Å^2; medium proteins those with surface areas between 10,000 and 17,000 Å^2; and large proteins, those with surface areas greater than 17,000 Å^2. Within these three subsets, the fractions of interface residues were 35%, 23%, and 19% respectively. The effect of this division is presented later.

A direct comparison between the previous and the updated data sets were made by applying the same neural network architecture and the same protocol as in our previous study on the new training set of 1156 chains. In particular, within the training set, a surface residue previously was counted as in the interface only if it had at least three, not one, interfacial contacts. The choice of a threshold of three was made in the previous study to make a less number of more reliable predictions. The new test set of 100 chains had a total of 14,678 surface residues, of which 3685 (or, 25%) had at least one interfacial contact (Table I). When the previous predictor was applied on the new test set, 2877 predictions of interface residues were

TABLE I. Statistics of Data Set

	Training set			Test set		
Group	Chains	Surface res.	Interface res.	Chains	Surface res.	Interface res.
Small protein	387	34046	11936 (35%)	47	3963	1391 (35%)
Medium protein	477	83636	19314 (23%)	35	5789	1356 (23%)
Large protein	292	92877	17688 (19%)	18	4926	938 (19%)
Homodimer	756	141818	30993 (22%)	42	6737	1657 (25%)
Heterdimer	400	68741	17945 (26%)	58	7941	2028 (26%)
All	1156	210559	48938 (23%)	100	14678	3685 (25%)

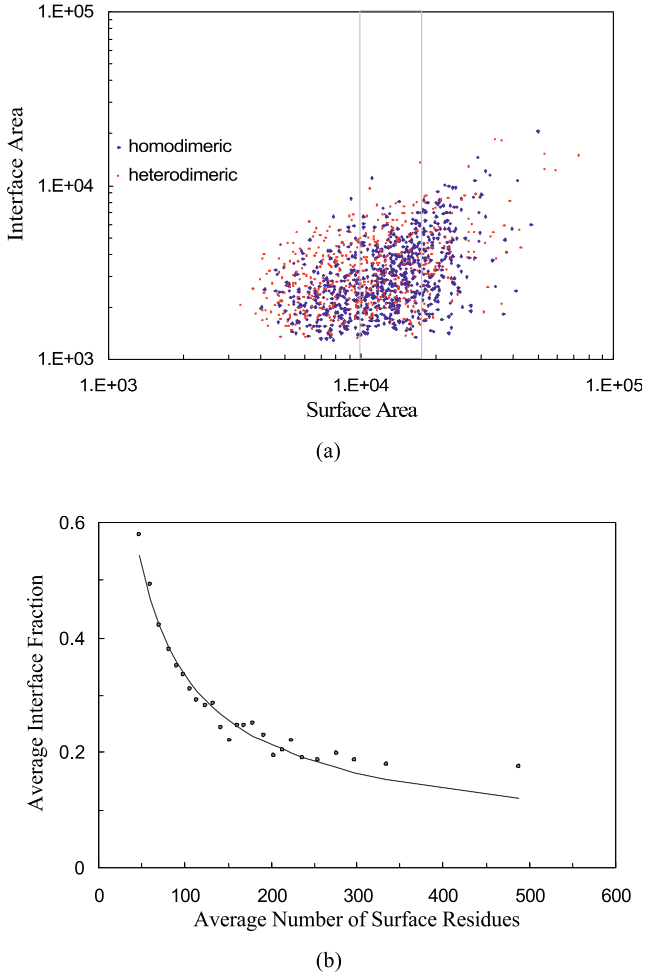

Fig. 2. Interface size relative to total protein surface. (**a**) Scatter plot of interface area versus total surface area. Areas are in Å^2. Homodimeric and heterodimeric chains are plotted in different colors. The interval in total surface area, between 10,000 and 17,000 Å^2, defining medium proteins is shaded. (**b**) Fraction of interface residues versus total number of surface residues. After sorting the 1256 chains in the data set in according to the number of surface residues, averages in batches of 50 chains were calculated for this quantity and the fraction of interface residues.

made, with 73.7% accuracy and 42% coverage of native interface residues. These results were likely inflated because 55 of the 100 test proteins had higher than 30% sequence identities to the old training set. In comparison, the newly trained predictor made 915 predictions, with an accuracy of 85.7% but coverage of just 17% (Table II).

Refinement of PPISP

Despite the significantly increased accuracy, the low coverage obtained with the old protocol is not desirable. A better compromise between accuracy and coverage had to be found. Our goal was to increase the coverage to ~50% and achieve as high an accuracy as possible. The first obvious thing to do was to lower the threshold for designating interface residues in the training set, to a minimum of just one interfacial contact. This raised the population of interface residues from 19% to 23% and had the desired effect. The test results showed an 81.6% accuracy with 28% coverage. Therefore we kept the threshold of one for interface residues in later developments, which are now presented in turn. The resulting changes in performance are summarized in Table II.

To examine the importance of including solvent accessibility in the input, we trained the neural network with just the sequence profile as input. The test results were 75.4% accurate with coverage of 36%. This was better than random prediction by 30 percentage points, and the inclusion of solvent accessibility amounted to an increase of six percentage points in accuracy. Since solvent accessibilities of different residues may have random fluctuations, we also tested the use of a smoothed solvent accessibility, obtained as the average over the residue under consideration and its six spatially nearest neighbors. This further increased the accuracy from 81.6% to 83.2%, with some deterioration in coverage (changing from 28% to 23%). The use of average solvent accessibility was kept.

The window size can affect performance. Too small a window size means that only a few spatial neighbors are included, which may not give a strong enough signal for accurate prediction. On the other hand, with too large a window size, distant neighbors are included, which may introduce noise. To optimize performance, window sizes between 10 and 20 were tested, and a size of 15 gave the best results, with accuracy at 84.5 and coverage at 22%. This size was retained.

Since we only focused on a single interface for each protein in our data set, for proteins that form trimers and higher oligomers, residues found in interfaces other than the selected one were classified as noninterface. We wondered whether these residues could mislead the neural network to some extent because they should have the characteristics of interface. Residues in these other interfaces accounted for 5% of the surface residues. When these residues were eliminated from the training set (reducing

TABLE II. Performance of Various Training Sets and Protocols

Test	Training	Protocol	(Strict) accuracy[c] (%)	Coverage (%)
All	All	Random	(26.1) 45.1	21
All	All	thrd = 3, w = 20, SA[a]	(69.4) 85.7	17
All	All	thrd = 1, w = 20, SA	(61.8) 81.6	28
All	All	thrd = 1, w = 20, no SA	(55.4) 75.4	36
All	All	thrd = 1, w = 20, avgSA	(64.2) 83.2	23
All	All	thrd = 1, w = 15, avgSA	(64.6) 84.5	22
All	All	thrd = 1, w = 15, avgSA, intfM[b]	(62.8) 86.5	26
All	All	thrd = 1, w = 15, avgSA, intfM, trim1/4	(56.6) 82.3	38
All	All	thrd = 1, w = 15, avgSA, intfM, trim1/3[b]	(52.8) 78.6	51
All	All	thrd = 1, w = 15, avgSA, intfM, trim1/2	(48.3) 74.6	61
Small	All	thrd = 1, w = 15, avgSA, intfM, trim1/3	(60.8) 91.7	56
Small	Small	thrd = 1, w = 15, avgSA, intfM, trim1/3	(61.7) 92.2	50
Medium	All	thrd = 1, w = 15, avgSA, intfM, trim1/3	(47.8) 74.5	51
Medium	Medium	thrd = 1, w = 15, avgSA, intfM, trim1/3	(43.2) 69.7	50
Large	All	thrd = 1, w = 15, avgSA, intfM, trim1/3	(49.1) 65.9	43
Large	Large	thrd = 1, w = 15, avgSA, intfM, trim1/3	(52.6) 72.1	45
Homodimer	All	thrd = 1, w = 15, avgSA, intfM, trim1/3	(57.2) 82.0	54
Homodimer	Homodimer	thrd = 1, w = 15, avgSA, intfM, trim1/3	(57.6) 82.2	54
Heterodimer	All	thrd = 1, w = 15, avgSA, intfM, trim1/3	(49.2) 75.9	48
Heterodimer	Heterodimer	thrd = 1, w = 15, avgSA, intfM, trim1/3[b]	(48.4) 77.2	49
Heterodimer	All	thrd = 1, w = 15, avgSA, intfM	(55.1) 81.6	32
Heterodimer	Heterodimer	thrd = 1, w = 15, avgSA, intfM[b]	(55.0) 81.8	36
Small heter.	Small heterodimer	thrd = 1, w = 15, avgSA, intfM[b]	(61.1) 88.2	52
Small heter.	Small heterodimer	thrd = 1, w = 15, avgSA[b]	(63.0) 85.2	48

[a] The original PPISP protocol. "SA" means inclusion of solvent accessibilities of individual residues as input, "no SA" means that solvent accessibilities were not included, and "avgSA" means that solvent accessibilities were included as averages over the residue under consideration and its six spatially nearest neighbors.

[b] These six entries are neural network models used to build cons-PPISP. "intfM" means that, for chains forming trimers or higher oligomers, residues in interfaces other than the largest one were eliminated from the training set. "trim1/3" means that one third of noninterface surface residues were randomly trimmed from the training set.

[c] Strict accuracy, given in parentheses, was the percentage of native interface residues among all predictions. In comparison, accuracy was calculated by counting the four nearest spatial neighbors of a native interface residue also as correct predictions.

the collection of noninterface residues from 161,621 to 150,649), both the accuracy and coverage were increased. The former went from 84.5% to 86.5%, and the latter went from 22% to 26%. While the improvement in accuracy may be attributed to rectification of residue mis-classification, the improvement in coverage perhaps largely arose from the mere decrease in the number of noninterface residues, making the training set more balanced between the interface and noninterface collections.

The issue of imbalance between the interface and noninterface collections in the training set was further investigated in order to extend the improvement on the coverage of native interface residues. Our suspicion was that the imbalance caused low predictions of interface residues. Therefore we randomly removed some of the noninterface residues to obtain a more balanced training set. The results of randomly trimming one quarter, one third, and one half of noninterface residues are shown in Table II. As suspected, the trimming led to significant improvement in coverage, with deeper trimming giving higher coverage but correspondingly lower accuracy. A good compromise appeared to be obtained with the one-third trimming, which increased the interface population to 33% in the training set. The coverage now increased to 51%, with the accuracy stood at 78.6%.

As already noted, the obvious decrease in the interface fraction of surface residues with increasing protein surface area motivated us to divide the data set into three sizes. Training and testing were done for each size separately. In comparison to training with the full set, training with the subset of small proteins led to a modest increase in accuracy, from 91.7% to 92.2%, for small proteins in the test set, but a significant decrease in coverage, from 56% to 50%. Training with the subset of medium proteins led to deteriorations in both accuracy (from 74.5% to 69.7%) and coverage (from 51% to 50%) for medium proteins in the test set. However, training with the subset of large proteins appeared advantageous for large proteins in the test set, with accuracy increasing from 65.9% to 72.1% and coverage increasing from 43% and 45%. With the protocol of training with the full set for small and medium proteins but with the subset of large protein for large proteins, a total of 3,529 interface residues were predicted among the 14,678 surface residues of the test set, with an accuracy of 80.2% and coverage of 51%.

There have been some indications that the interface characteristics are different between homodimers and heterodimers.[10,19,20,29] For homodimers that are formed concomitantly with the folding of their subunits, the interfaces are essentially the same as the interiors of

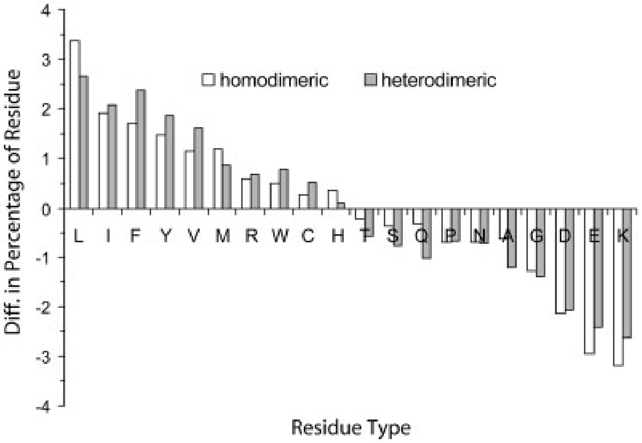

Fig. 3. Difference in residue percentage between the interface and noninterface collections in the training set. Results for homodimeric and heterodimeric chains are shown in different bars to show potential differences between the two kinds of chains.

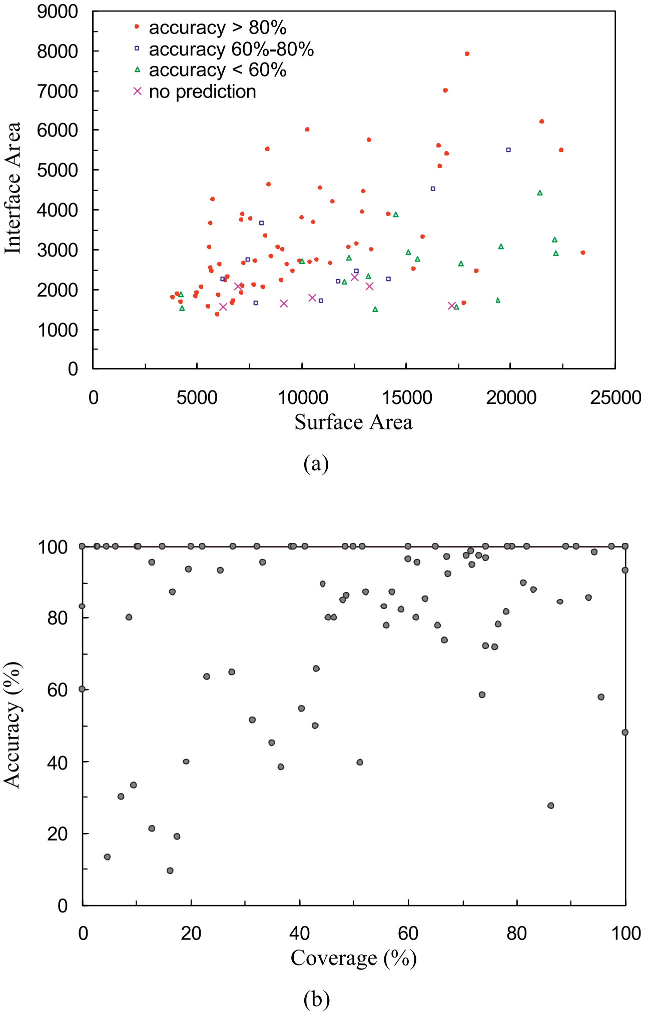

Fig. 4. Prediction accuracy and coverage for the test set of 100 chains, with small and medium proteins predicted from the full training set and large proteins predicted from the group of large protein in the training set. The overall prediction accuracy was 80.2% with 51% coverage. (**a**) Scatter plot of interface area versus total surface area, grouped according to prediction accuracy. Areas are in Å^2. (**b**) Scatter plot of accuracy versus coverage.

individually folded proteins, and thus should be enriched in nonpolar residues and deficient in charged residues. On the other hand, for a protein that is a functioning entity both before and after forming a complex with a different protein, the interface residues are expected to have characteristics intermediate between protein interiors and noninterface portions of protein surfaces. In our data set, we did find some qualitative difference in amino acid composition between homodimeric and heterodimeric interfaces (Fig. 3). Charged residues indeed showed higher propensities in heterodimeric interfaces, but nonpolar residues exhibited mixed behavior. We separately trained and tested homodimers and heterodimers in our data set. In comparison to training with the full set, training with only homodimers did not show any improvement for homodimers in the test set, but training with only heterodimers gave slightly better results for heterodimers in the test set. The indifference for homodimers perhaps can be attributed to the overpopulation of homodimers in the full training set. Improvement for heterodimers was seen regardless of whether random trimming of noninterface residues was introduced in the training set. The potential of improving predictions of heterodimeric interfaces by separately training with heterodimers will become increasingly important as the population of heterodimers grows in the PDB.

For each of the neural network models, the result presented in Table II was typically from the round of training giving the highest accuracy.

Comparison with other methods is difficult because of the variety of data sets used and the difference in definitions of interface and surface residues. In particular, Koike and Takagi[31] used a support vector machine to predict interaction sites using definitions of interface and surface similar to ours. The strict accuracy and coverage of their interface prediction were 54–56% and 50%, respectively, with all interfaces in multimers counted. When all interfaces were included, the strict accuracy and coverage of our predictions were 61% and 48%, respectively. Fariselli et al.[29] used neural network to predict interaction sites and reported strict accuracy and coverage of 72% and 56%, respectively, for a selected data set of 226 heterodimers. However, their interface residues were defined with a C_α cutoff distance of 12 Å. This definition includes far more surface residues as interface sites (40% relative to the 23% in the present study), and significantly decreases the difficulty of interface prediction. For example, random predictions would have expected strict accuracy at 40% according to their interface–residue definition but only 23% in our study.

Consensus Approach

With the division of the data set into three subsets according to surface size, overall accuracy of 80.2% and coverage of 51% in the test set were achieved. However, prediction accuracy and coverage were very uneven, as Figure 4 shows. Of the 100 proteins in the test set, 43 were

TABLE III. Comparison of Predictions by Best Individual Neural Network Models and the Consensus Approach

				Best Model			cons-PPISP		
Group	Complexes	Unique chains	Real interface residues	Pred. interface residues	(Strict) accuracy (%)	Coverage (%)	Predicted interface residues	(Strict) accuracy (%)	Coverage (%)
Enzyme-inhibitor	22	33	657	619	(47.3) 67.7	45	661	(49.9) 70.8	50
Other	11	21	398	349	(25.2) 40.1	22	358	(31.0) 48.0	28
Difficult	7	14	367	397	(30.2) 49.4	33	276	(37.3) 56.2	28
All	40	68	1422	1365	(36.7) 55.3	35	1295	(42.0) 61.4	38

quite successful with accuracy ≥ 70% and coverage ≥ 50%, and 12 were reasonably successful with accuracy ≥ 80% and coverage ≥ 30%. Among the remaining 45 proteins, five had coverage ≥ 50% but accuracy ≤ 60%, indicating over-prediction, and 17 had accuracy ≥ 80% but coverage ≤30%, indicating under-prediction, seven had no predictions at all, and 16 had sufficient predictions, but these were not accurate.

The problem of over-prediction and under-prediction was universal with all the neural network models, though the balance between the two trends varied. A model with higher accuracy tended to have more severe under-prediction, whereas a model with higher coverage tended to have more severe over-prediction. Another problem with an individual neural network is the decision on how many rounds of training to carry out. In general, prediction accuracy improves in the beginning rounds and then deteriorates in later rounds, while coverage decreases monotonically. Thus under-training may bring benefit to coverage at the expense of accuracy.

We sought to tackle the problem of over-prediction and under-prediction by combining results from models that covered a range of accuracy and coverage. Predictions from the different models were clustered according to spatial relations, and the cluster or clusters of residues predicted with the highest consensus were taken as the final predictions of interface residues. Specifically, for each surface residue a consensus score was defined as the number of models predicting that residue as in the interface. All surface residues were then ranked by their consensus scores. Clusters were grown from the highest ranked residues. If a large number of predictions were made with high consensus scores, collection of interface residues was limited to these predictions, otherwise the process was extended to predictions with lower consensus scores. Further details of the clustering and collection procedure are given in Materials and Methods. This consensus approach was able to bring satisfactory solutions to the 22 proteins that had over-prediction or under-prediction problems.

In the end we settled on a suite of six types of neural network models (Table II) to include in the consensus approach: (1) “intfM” trained up to nine rounds on the full test set; (2) “intfM” trained up to 10 rounds on the subset of heterodimers; (3) “intfM” trained up to 10 rounds on the subset of small heterodimers; (4) “intfM, trim1/3” trained up to nine rounds on the full test set; (5) “intfM, trim1/3” trained up to 10 rounds on the subset of heterodimers; and (6) “all intf” trained from rounds 2 to 21 on the subset of small heterodimers. The different rounds of training on the six types of neural networks made up a total of 68 predictors. For easy reference, the consensus approach was named cons-PPISP.

The six types of predictors were selected according to their accuracy and coverage for different types of test proteins (homodimer versus heterodimer; small, medium, or large size). Different considerations led to the elimination of a number of predictors. For example, the predictor trained with the subset of homodimers was not included because such training did not show improvement for homodimers over training with the full set. Though separate training with the subset of large proteins modestly improved prediction, that predictor was not included because of the sparsity of large proteins in the data set and the possibility that many large proteins can be divided into domains of smaller sizes. We did include predictors trained with the subset of small heterodimers because such training led to a moderate increase in accuracy, though with a corresponding decrease in coverage, for small proteins. In this way the selection for predictors was narrowed to those trained with the full set, the subset of all heterodimers, and the subset of small heterodimers. For each training set, two predictors were selected: one with the highest accuracy (with corresponding low coverage) and one with the highest accuracy at 50% coverage. The resulting six types of predictors, coupled with the freedom of successive rounds of training with varying accuracy and coverage, afforded a uniform span of a wide range of coverage (from 20% to 70%), which in turn allowed for the gradual growth of clusters from the most confident to the least confident predictions.

Test on a Benchmark Set for Protein–Protein Docking

Chen et al.[35] collected a set of protein–protein complexes as a docking benchmark, which included 22 enzyme-inhibitor pairs, 11 other kinds of pairs, and seven pairs that were deemed difficult for docking methods. We tested cons-PPISP on the 68 unique chains of these 40 complexes (Table III). There were also 19 antibody-antigen pairs in the benchmark, but in our data set for training and testing the neural networks antibody–antigen interfaces were filtered out. As noted by Jones and Thornton,[8] “antibody–protein interactions are relatively ”happenstance“ and are selected principally by the strength of the binding constant, without being subject to evolutionary optimization over many years.” Indeed many different parts of a protein

TABLE IV. Prediction Results for the NMR Set

Complex[a]	PDB[b]	Sequence length	Real interface residues[c]		Predicted interface residues	(Strict) accuracy (%)	Coverage (%)
			CSP	X-ray			
Cytochrome b5/cytochrome c[43]	1cyo	88	17		21	(38.1) 76.2	47
Thrombomodulin/thrombin[37,44]	1dqbA	83	32	14 1dx5I/M	20	(70.0) 95.0	44
p47 C-term./p97 N-term.[40,45]	1i42A	89	23	19 1s3sG/F	21	(57.1) 95.2	52
Troponin C/troponin I[39,46,48,49]	1ncx	162	18	22 1a2xA/B	20	(40.0) 85.0	44
HP1 chromo/histone H3[41,42,66]	1ap0	73	10	12 1knaA/P	26	(15.4) 38.5	40
B domain of protein A/IgG[38,51]	1bdd	60	14	12 1fc2C/D	23	(43.5) 65.2	71
UBC9/SUMO-1[67,68]	1u9aA	159	19		20	(30.0) 75.0	32
	1a5r	103	11		23	(21.7) 34.8	46
All			144		174	(38.5) 69.0	47

[a] The first of the pair of proteins was the prediction target. For UBC9/SUMO-1, both partners were studied, with the top entries referring to UBC9 and the bottom entries referring to SUMO-1.

[b] PDB code for unbound protein used in interface prediction.

[c] If a residue is identified by chemical shift perturbation (CSP) as in an interface but is buried in the unbound structure, that residue was not counted as an interface residue. PDB code for a protein complex and chain identities used for obtaining interface residues are listed below the number of interface residues.

surface, as found for lysozyme and other proteins, can be targets for monoclonal antibodies. Methods like ours that heavily rely on evolutionary information are not particularly suited for predicting antibody–antigen interfaces.

For the 33 chains forming enzyme-inhibitor complexes, the "intfM, trim1/3" model trained eight rounds on the subset of heterodimers in our training set gave the best performance among all the neural network models, with accuracy at 67.7% and coverage of 45%. For the 21 chains forming other complexes, the best performance was obtained with the "intfM, trim1/3" model trained five rounds on the full training set, with accuracy at 40.1% and coverage of 22%. For the "difficult" set of 14 chains, the "intfM, trim1/3" model trained four rounds on the subset of heterodimers gave the best performance, with accuracy at 49.4% and coverage of 33%. Altogether, these separately best-performing models had an accuracy of 55.3% and coverage of 35% for the 68 proteins in the benchmark set.

cons-PPISP outperformed the individual neural network models for all the three subsets of proteins. For the enzyme-inhibitor subset, cons-PPISP increased prediction accuracy by three percentage points to 70.8% and coverage by five percentage points to 50%. For the "other" subset, cons-PPISP increased prediction accuracy by eight percentage points to 48.0% and coverage by six percentage points to 28%. For the "difficult" subset, cons-PPISP increased prediction accuracy by 7 percentage points to 56.2%, though at the expense of decreasing coverage to 28%. Altogether, the prediction accuracy for the 68 proteins was increased by cons-PPISP to 61.4% and coverage increased to 38%. What makes this enhancement in performance all the more important is that it was obtained without having to manually choose which neural network model to use and how many rounds of training to do.

These results were obtained by using the unbound structures of the complex-forming proteins. The performance of cons-PPISP using the unbound structures was nearly the same as using the bound structures. In the latter case the overall prediction accuracy for the 68 proteins was 63.6% and coverage was 39%. Such robustness with respect to unbound structures was designed into PPISP and was already demonstrated in our previous study.[7]

Test on a Set of NMR-Characterized Proteins

To further demonstrate the predictive power of cons-PPISP, we collected from the literature proteins whose interfaces have been characterized by NMR chemical shift perturbation. Those with sequence homologies with the training set were removed, resulting in a total of eight proteins (Table IV). Three of these, thrombomodulin, troponin C, and the B domain of protein A, had X-ray structures for their respective complexes at the time. These X-ray structures were not selected into our data set for training neural networks for two different reasons: (1) less than 20 contact residues for the complexes between thrombomodulin and thrombin[37] and between the B domain of protein A and the Fc fragment of immunoglobulin G (IgG);[38] (2) fewer than 40 residues in the partner chain, troponin I, in the case of troponin C.[39] X-ray and NMR structures of complexes formed by two other proteins, the adaptor protein p47 with the AAA ATPase protein p97[40] and the HP1 chromo domain with histone H3,[41,42] have recently been determined. The remaining three still do not have X-ray structures for their complexes. According to chemical shift perturbation, the eight proteins have a total of 144 interface residues. Cons-PPISP made 174 predictions, with an accuracy of 69% and coverage of 47%.

We now present detailed comparisons of the cons-PPISP predictions with experiments. The interface formed by cytochrome b_5 with cytochrome *c* has recently been reexamined by Shao et al.[43] Their chemical shift perturbation results suggest that the following 17 residues of cytochrome b_5 made up the interface with cytochrome *c*: K34, E37, H39, G41, E44-L46, G52, T55, E56, E59-V61, H63, S64, A67, and L70 [Fig. 5(a)]. The last eight of these were among the 21 predicted interface residues. Eight other predicted residues (N57, F58, G62, D66, R68, E69, S71,

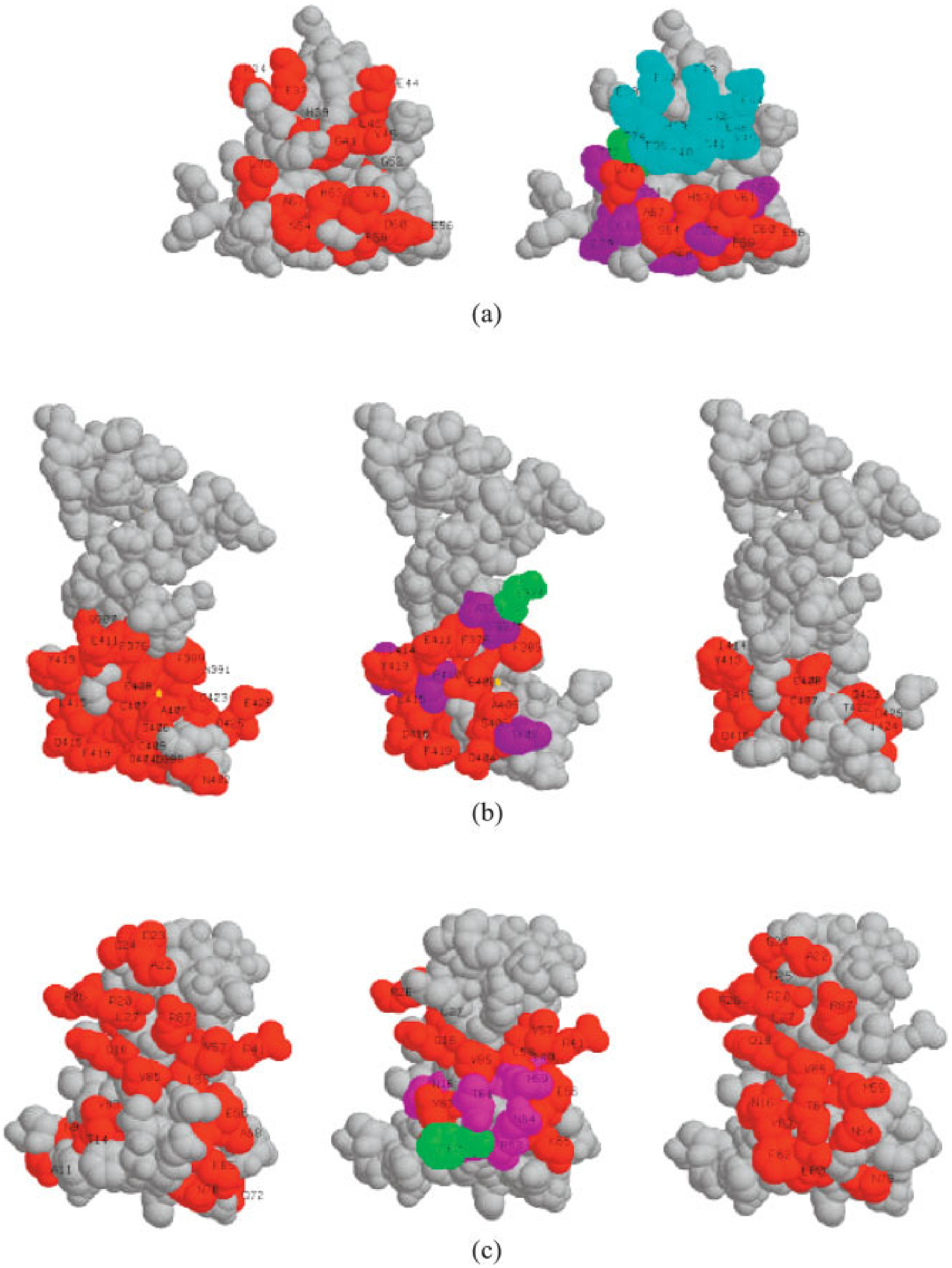

Fig. 5. Comparison of predicted interface residues with those identified by chemical shift perturbation or found in X-ray structures of protein complexes. (**a**-**h**) show the eight proteins in Table IV, in the order listed there. For each protein, NMR results, predictions, and X-ray results (if available) are displayed in the first, second, and third panels, respectively. Residues identified by NMR, X-ray, or predicted correctly are in red. Loosely correct predictions and false positives are in purple and green, respectively. In (**a**) and (**d**), a cluster of residues that was eliminated by cons-PPISP and appears to be within the interface is displayed in cyan.

and T73) were among the nearest neighbors of the interface sites suggested by NMR. The remaining five predictions (K5, D31, and F74-I76) were deemed false. Interestingly, another cluster of 11 residues with high consensus scores were eliminated by cons-PPISP because of the relatively smaller cluster size. These residues (F35 and E37-L46) turned out to cover the first half of the NMR-deduced interface sites. When these were included, the prediction accuracy increased from 76.2% to 84.4% while coverage increased from 47% to 82%.

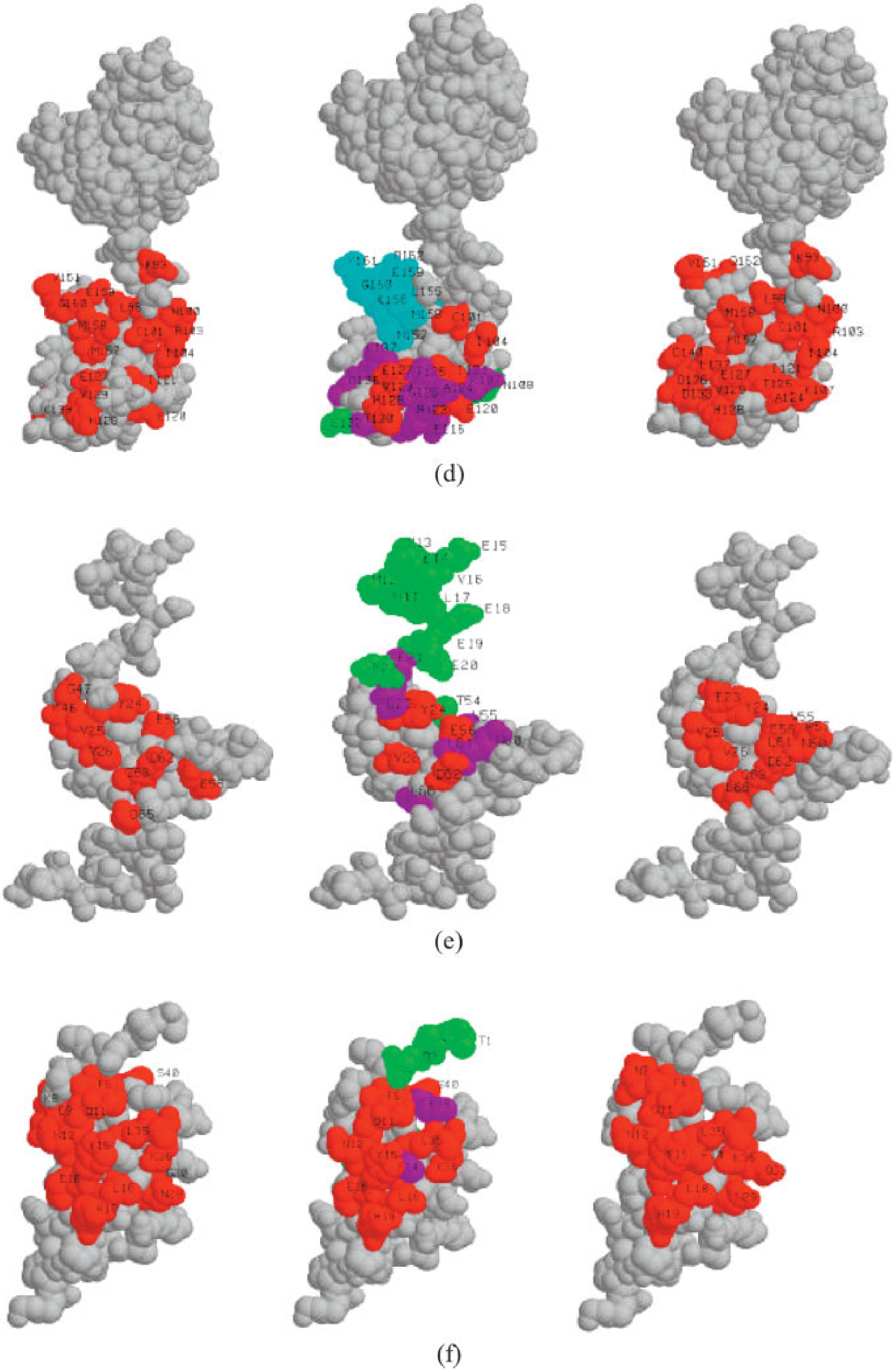

Figure 5. (Continued.)

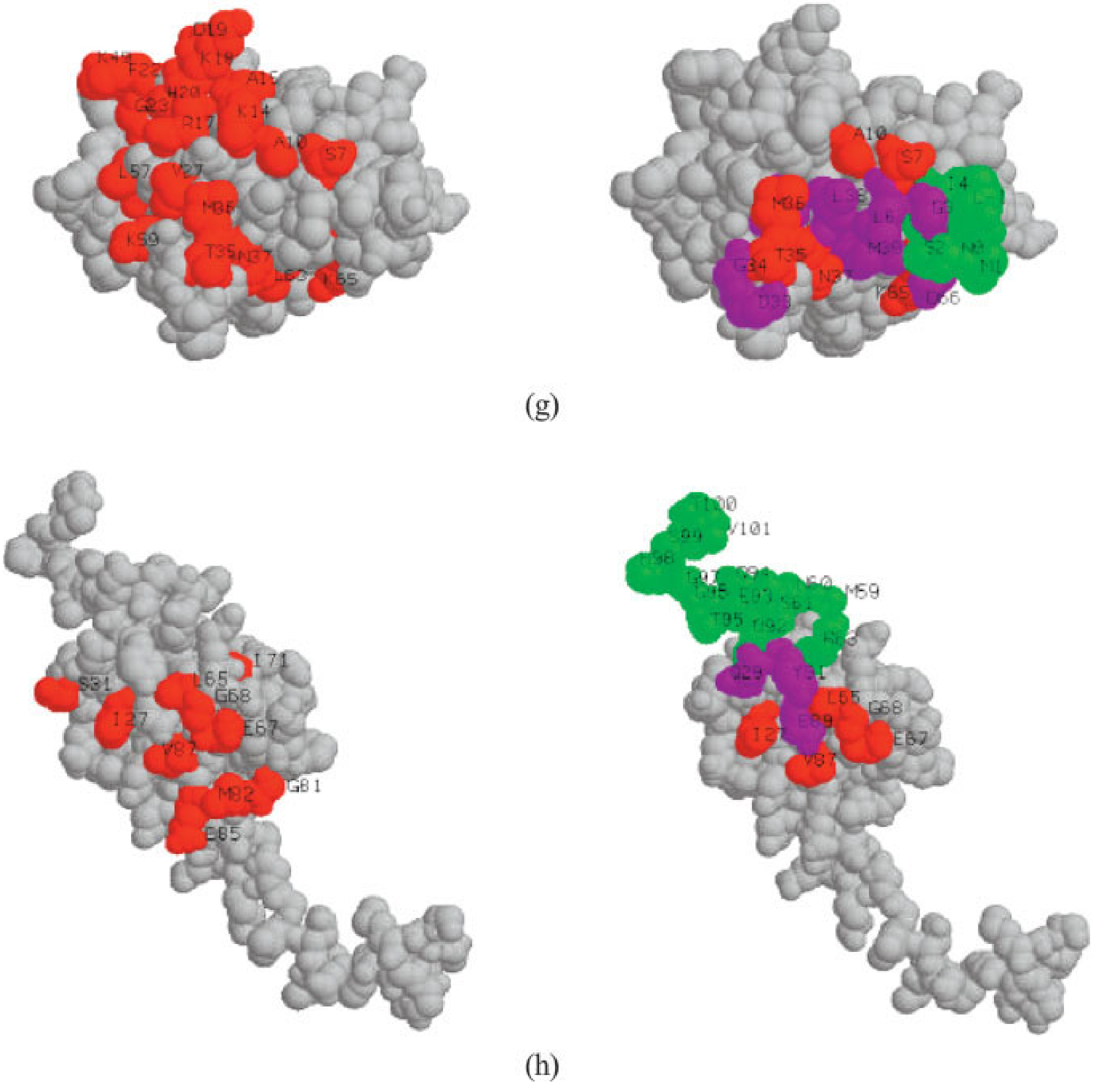

Figure 5. (Continued.)

Wood et al.[44] used chemical shift perturbation to characterized the interface formed by two epidermal growth factor-like domains (TMEGF45) of thrombomodulin with thrombin and identified the following 32 interface residues: F376, A377, Q387-Q392, A394, A397-D400, N402, Q404-C409, E411-I420, D423, D4256, and E426 [Fig. 5(b)]. Of these, 14 were among the 20 predicted interface residues. These correctly predicted residues were: F376, F389, Q404-S406, E408, C409, E411-Y413, L415, D416, F419, and I420. Five other predicted residues (A373, G375, T403, P410, and I414) were among the nearest neighbors of the NMR-deduced interface sites. Only a single prediction, E374, was deemed false by our criteria, though it was very close to the interface site F376.

Yuan et al.[45] identified the interaction surface of the C-terminal domain of the adaptor protein p47 with the AAA ATPase protein p97 by chemical shift perturbation, and obtained the following 23 interface residues: N9, A11, T14, Q18, R20, A22-G24, R26, L27, R41, V57, L58, K65-A68, Q72, N78, A82, V83, V85, and R87 [Fig. 5(c)]. Of these, 12 (Q18, R26, L27, R41, V57, L58, K65-L67, A82, V83, and V85) were among the 21 predicted interface residues. Eight other predicted residues (N16, I40, F43, M59, T61, P63, N64, and N81) were among the nearest neighbors of the NMR-deduced interface sites. The remaining predicted residue, F62, was near the NMR-deduced interaction sites but was nonetheless deemed false.

As shown in the recent structure of the ternary complex, troponin C forms four binding sites with three regions of troponin I as well as a coiled-coil formed by troponin I and troponin T.[46] We had no way of knowing which binding sites would be predicted. When the predicted residues were examined, they were found to form the binding surface for the N-terminal region of troponin I [Fig. 5(d)]. Interestingly, the binding affinity of the N-terminal peptide of troponin I has been found to be orders of magnitude higher than those of the inhibitory and C-terminal peptides.[47] The binding surface on troponin C for the N-terminal region of troponin I has been mapped by chemical shift perturbation,[48,49] and the following 18 interface residues were identified: K93, L98, N100, C101, R103, I104, E120-L122, E127-V129, K139, and M157-V161. Of these, eight (C101, I104, E120-L122, and E127-V129) were among the 20 predicted interface residues. Nine other predicted residues (K107, E116, R123-G126, T130, D136, and L137) were among the nearest neighbors of the NMR-deduced interface sites. The remaining three predictions, N108, I115, and E132, though in the vicinities of the NMR-deduced interface sites, were deemed false by our criteria. A second cluster of ten predicted residues with high consensus scores turned out to cover the remaining part of the binding surface for the N-terminal region of troponin I. When these residues (F151, D152, and L155-Q162) were included, the coverage of NMR-deduced interface sites increased from 44% to 72% while prediction accuracy remained high (changing from 85.0% to 83.3%).

Jacobs et al.[50] identified the histone H3-binding surface on the HP1 chromo domain by chemical shift perturbation and found the following ten interface residues: Y24-V26, K46, G47, E56, E58, D62, C63, and D65 [Fig. 5(e)]. Of these, four (Y24, V26, E56, and D62) were among the 26 predicted interface residues. Six other predicted residues (E21, E23, W55, N60, L61, and L66) were among the nearest neighbors of the NMR-deduced interface sites. Of the 16 predictions that were deemed false, 11 were from the N-terminal (H11-E20 and E22), and the other five were L30, D31, L42, T54, and F70.

Takahashi et al.[51] mapped the interaction surface of the B domain of protein A with the Fc fragment of IgG by chemical shift perturbation, and identified the following 14 interface residues: F6, K8, E9, Q11, N12, Y15, E16, L18, H19, N29, G30, L35, K36, and S40 [Fig. 5(f)]. Ten of these coincided with the predicted interface residues. These were: F6, Q11, N12, Y15, E16, L18, H19, L35, K36, and S40. Five other predicted residues (F14, F31, P39, Q41, and A43) were among the nearest neighbors of the NMR-deduced interface sites. The remaining eight of the 23 predicted residues, T1-N4 and N44-A47, were deemed false-positives. Though our method was not designed for antigen–antibody complexes, the interface predictions overall were actually satisfactory. Perhaps the implicated surface is an “easy” target for antibody.

Chen and coworkers[52,53] identified the interaction surfaces of the ubiquitin conjugation enzyme Ubc9 and small ubiquitin modifier 1 (SUMO-1) by chemical shift perturbation. On the Ubc9 side, the following 19 residues were obtained: S7, A10, K14, A15, R17-H20, F22, G23, V27, T35-N37, K49, L57, K59, L63, and K65 [Fig. 5(g)]. Six of these (S7, A10, T35-N37, and K65) were among the 20 predicted interface residues. Nine other predicted residues (G3, L6, P28, D33, G34, L38, M39, D66, and D67) were among the nearest neighbors of the NMR-deduced interface sites. The remaining five predicted residues, located at the N-terminal, were deemed false-positives. On the SUMO-1 side, the following 11 residues were identified by NMR as interaction sites: I27, S31, V38, L65, E67, G68, I71, G81, M82, E85, and V87 [Fig. 5(h)]. Five of these (I27, L65, E67, G68, and V87) were among the 23 predicted interface residues. Three other predicted residues (Q29, E89, and Y91) were among the nearest neighbors of the NMR-deduced interface sites. The remaining 15 predicted residues, V57, M59-S61, R63, Q92-V101, were located in the periphery of the NMR-deduced interaction sites, but were deemed false-positives.

For the five proteins with known structures for their complexes, comparison of the predicted interface residues with the native interfaces was also very encouraging. The high accuracy and coverage against NMR data were maintained for thrombomodulin, adaptor protein p47, troponin C, and protein A. For the HP1 chromo domain [Fig. 5(e)], when compared against the X-ray structure for the complex with the K9-methylated histone H3 tail,[41] prediction accuracy increased from 38.5% to 53.8% and coverage increased from 40% to 75%. These results suggest that cons-PPISP has the potential of complementing NMR and other experimental techniques in characterizing protein–protein interfaces.

From Dimer to Trimer

This work has been focused on protein dimers. However, a higher oligomer can almost always be constructed sequentially one dimeric interface at a time. Often the biological process actually follows such sequential steps. Such is the case for the activation of protein C. This process is initiated by the binding of thrombomodulin to thrombin, and the resulting complex then presents a suitable binding surface for protein C.[37] It appears possible to extend our method to predict a binding surface that is located on a protein complex instead of a single protein. The two protein chains in the complex can be treated as a single protein for the purpose of calculating solvent accessibility and neighbor lists. However, for the purpose of generating sequence profiles by PSI-BLAST, more robust results may be obtained by searching for sequence alignments separately for the two chains and then concatenating the two alignments.

Given the basic architecture of protein functioning machineries in the form of high oligomers, binding surfaces located on protein complexes certainly warrant systematic studies. Here we give preliminary results of cons-PPISP predictions for the protein C-binding surface on the thrombin–thrombomodulin complex. A cluster of 18 residues on the fourth epidermal growth factor-like domain of thrombomodulin was identified by cons-PPISP, though no residues from thrombin were predicted with high confidence. TMEGF4 has indeed been established as the site for binding protein C by mutational studies,[54–57] and indeed Fuentes-Prior et al.[37] constructed a structural model for the ternary complex with protein C contacting thrombin and TMEGF4. The 18 predicted residues were V345-A354 and C360-S367. These predictions appear consistent with residues that have been implicated being important for binding protein C, which include D349, E357, Y358, and F376.[54–57]

Test on CAPRI Targets

Although the first attempt to computationally reconstitute a complex by docking two proteins together dates back to the late 1970s,[58] up to now, protein–protein docking has remained largely an academic exercise. Progress in this direction is hampered by two major problems: the conformational changes that usually accompany complex formation and the lack of scoring functions that can discriminate efficiently between the correct docking solution and many false positives.[34] Identification of interface residues by a method like cons-PPISP holds great potential to simplify the docking problem by either restricting the search in the six-dimensional translational-rotational space or eliminating false positives after the search. This potential was realized in a recent collaborative participation with the Bonvin group in the latest rounds (4 and 5) of CAPRI. In this collaboration, interface residues predicted by cons-PPISP as well as experimental mutagenesis and other types of data were used as ambiguous interaction restraints to drive the docking process, through the HADDOCK program developed by the Bonvin group.[6]

The results of our CAPRI participation will be reported elsewhere.[59] Here we give a short summary of the cons-PPISP predictions on the CAPRI 4 and 5 targets. For a total of 11 protein targets (no prediction was necessary for two antibodies), three (cohesin of T11/T12,[60] MYPT1 of T14,[61] and colicin D of T15[62]) had higher than 80% prediction accuracy, and four (TBE monomer of T10,[63] dockerin of T11/T12, immunity protein of T15, and the xylanase of T18[64]) had accuracy between 50% and 80%. Interface predictions were poor for two antigens [in T13 (still unpublished) and T19[65]] and two large proteins (PP-1 of T14 and TAXI of T18). These blind test results on the CAPRI targets are consistent with the results presented earlier on the other test sets.

CONCLUSIONS

We have developed a robust program, cons-PPISP, for predicting interface residues in protein–protein complexes. Taking advantage of the significant expansion of protein complexes in the Protein Data Bank, we carefully analyzed characteristics of proteins, provided rationalizations for these characteristics, and refined the neural network models to capture these characteristics. These ideas for model design may be useful for future developments as the PDB further expands. At the present state, cons-PPISP promises to complement NMR and other experimental techniques in characterizing protein–protein interfaces.

Ultimately three-dimensional structures of protein complexes may be indispensable for a full understanding of the mechanisms of protein–protein interactions and protein functions. So far only a small fraction of the hundreds and thousands of putative protein complexes have their structures determined. Many protein complexes are formed by weak interactions and may not be amenable to structure determination by X-ray or NMR. Thus computational methods that can build structural models for protein complexes will become more and more important as the number of individual protein structures rapidly grows. Incorporation of cons-PPISP predictions in a docking program in the latest CAPRI rounds has demonstrated its ability in guiding the docking process.

ACKNOWLEDGMENTS

This work was supported in part by NIH grant GM58187.

REFERENCES

1. Ito T, Chiba T, Osawa R, Yoshida M, Hattori M, Sasaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 2000;98:4569–4574.
2. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 2000;403:623–627.
3. Lichtarge O, Sowa ME, Philippi A. Evolutionary traces of functional surfaces along the G protein signaling pathway. Methods Enzymol 2001;344:536–556.
4. Sowa ME, He W, Slep KC, Kercher MA, Lichtarge O, Wensel TG. Prediction and confirmation of a site critical for effector regulation of RGS domain activity. Nat Struct Biol 2001;8:234–237.
5. Zhou H-X. Improving the understanding of human genetic disease through predictions of protein structures and protein-protein interaction sites. Curr Med Chem 2004;11:539–549.
6. Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 2003;125: 1731–1737.
7. Zhou H-X, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 2001;44:336–343.
8. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
9. Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches. J Mol Biol 1997;272:121–132.
10. Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 1997;272:133–143.
11. Conte LL, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177–2198.
12. Hu Z, Ma B, Wolfson H, Nussinov R. Conservation of polar residues as hot spots at protein interfaces. Proteins 2000;39:331–342.
13. Valdar WSJ, Thornton JM. Protein-protein interfaces: analysis of amino acid conservation in homodimers. Proteins 2001;42:108–124.
14. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N. Residue frequencies and pairing preferences at protein-protein interfaces. Proteins 2001;43:89–102.
15. Armon A, Graur D, Ben-Tal N. Consurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 2001;307:447–463.
16. Landgraf R, Xenarios L, Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 2001;307:1487–1502.
17. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 2002;18:S71–S77.
18. Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME, Lichtarge O. 2002. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J Mol Biol 316:139–154.
19. Lichtarge O, Sowa ME. Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol 2002;12:21–27.
20. Ofran Y, Rost B. Analysing six types of protein-protein interfaces. J Mol Biol 2003;325:377–387.
21. Nooren IMA, Thornton JM. Structural characterization and functional significance of transient protein-protein interactions. J Mol Biol 2003;325:991–1018.
22. Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kavraki L, et al. An accurate, sensitive, and scalable method to identify functional sites in protein structures. J Mol Biol 2003;326: 255–261.
23. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci USA 2003;100:5772–5777.

24. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25: 3389–3402.
25. Rost B, Sander C. Progress of 1D protein structure prediction at last. Proteins 1995;23:295–300.
26. Jones D. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
27. Shan Y, Wang G, Zhou H-X. Fold recognition and accurate query-template alignment by a combination of PSI-BLAST and threading. Proteins 2001;42:23–37.
28. Minakuchi Y, Satou K, Konagaya A, Ito T. Prediction of protein-protein interaction sites using support vector machines. Genome Informatics 2002;13:322–323.
29. Fariselli P, Pazos F, Valencia A, Casadio R. Prediction of protein-protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 2002;269:1356–1361.
30. Ofran Y, Rost B. Predicted protein-protein interaction sites from local sequence information. FEBS Lett 2003;544:236–239.
31. Koike A, Takagi T. Prediction of protein–protein interaction sites using support vector machines. Protein Eng Des Sel 2004;17:165–173.
32. 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;338:181–199.
33. Yan CH, Honavar V, Dobbs D. Identification of interface residues in protease-inhibitor and antigen-antibody complexes: a support vector machine approach. Neural Comput Appl 2004;13:123–129.
34. Janin J, Henrick K, Moult J, Eyck LT, Sternberg M, Vajda S, Vakser I, et al. CAPRI: a critical assessment of predicted interactions. Proteins 2003;52:2–9.
35. Chen R, Mintseris J, Janin J, Weng Z. A protein-protein docking benchmark. Proteins 2003;52:88–91.
36. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
37. Fuentes-Prior P, Iwanaga Y, Huber R, Pagila R, Rumennik G, Seto M, Morser J, Light DR, Bode W. Structural basis for the anticoagulant activity of the thrombin-thrombomodulin complex. Nature 2000;404:518–525.
38. Deisenhofer J. Crystallographic refinement and atomic models of a human Fc fragment and its complex with fragment B of protein A from Staphylococcus aureus at 2.9- and 2.8-A resolution. Biochemistry 1981;20:2361–2370.
39. Vassylyev DG, Takeda S, Wakatsuki S, Maeda K, Maeda Y. Crystal structure of troponin C in complex with troponin I fragment at 2.3-Å resolution. Proc Natl Acad Sci USA 1998;95: 4847-4852.
40. Dreveny I, Kondo H, Uchiyama K, Shaw A, Zhang X, Freemont PS. Structural basis of the interaction between the AAA ATPase p97/VCP and its adaptor protein p47. EMBO J 2004;23:1030-1039.
41. Jacobs SA, Khorasanizadeh S. Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. Science 2002;295: 2080-2083.
42. Nielsen PR, Nietlispach D, Mott HR, Callaghan J, Bannister A, Kouzarides T, Murzin AG, Murzina NV, Laue ED. Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. Nature 2002;416:103-107.
43. Shao W, Im SC, Zuiderweg ER, Waskell L. Mapping the binding interface of the cytochrome b5-cytochrome c complex by nuclear magnetic resonance. Biochemistry 2003;42:14774–14784.
44. Wood MJ, Benitez BAS, Komives EA. Solution structure of the smallest cofactor-active fragment of thrombomodulin. Nat Struct Biol 2000;7:200–204.
45. Yuan X, Shaw A, Zhang X, Kondo H, Lally J, Freemont PS, Matthews S. Solution structure and interaction surface of the C-terminal domain from p47: a major p97-cofactor involved in SNARE disassembly. J Mol Biol 2001;311:255–263.
46. Takeda S, Yamashita A, Maeda K, Maeda Y. Structure of the core domain of human cardiac troponin in the Ca(2+)-saturated form. Nature 2003;424:35–41.
47. Li MX, Saude EJ, Wang X, Pearlstone JR, Smillie LB, Sykes BD. Kinetic studies of calcium and cardiac troponin I peptide binding to human cardiac troponin C using NMR spectroscopy. Eur Biophys J 2002;31:245–256.
48. Gasmi-Seabrook GMC, Howarth JW, Finley N, Abusamhadneh E, Gaponenko V, Brito RMM, Solaro RJ, Rosevear PR. Solution structures of the C-terminal domain of cardiac troponin C free and bound to the N-terminal domain of cardiac troponin I. Biochemistry 1999;38:8313–8322.
49. Mercier P, Li MX, Sykes BD. Role of the structural domain of troponin C in muscle regulation: NMR studies of Ca2+ binding and subsequent interactions with regions 1-40 and 96-115 of troponin I. Biochemistry 2000;39:2902–2911.
50. Jacobs SA, Taverna SD, Zhang Y, Briggs SD, Li J, Eissenberg JC, Allis CD, Khorasanizadeh S. Specificity of the HP1 chromo domain for the methylated N-terminus of histone H3. EMBO J 2001;20:5232–5241.
51. Takahashi H, Nakanishi T, Kami K, Arata Y, Shimada I. A novel NMR method for determining the interfaces of large protein–protein complexes. Nat Struct Biol 2000;7:220–223.
52. Liu Q, Jin C, Liao X, Shen Z, Chen DJ, Chen Y. The binding interface between an E2 (UBC9) and a ubiquitin homologue (UBL1). J Biol Chem 1999;274:16979–16987.
53. Tatham MH, Kim S, Yu B, Jaffray E, Song J, Zheng Z, Rodriguez MS, Hay RT, Chen Y. Role of an N-terminal site of Ubc9 in SUMO-1, -2, and -3 binding and conjugation. Biochemistry 2003; 42:9959–9969.
54. Zushi M, Gomi K, Honda G, Kondo S, Yamamoto S, Hayashi T, Suzuki K. Aspartic acid 349 in the fourth epidermal growth factor-like structure of human thrombomodulin plays a role in its Ca(2+)-mediated binding to protein C. J Biol Chem 1991;266: 19886–19889.
55. Nagashima M, Lundh E, Leonard JC, Morser J, Parkinson JF. Alanine-scanning mutagenesis of the epidermal growth factor-like domains of human thrombomodulin identifies critical residues for its cofactor activity. J Biol Chem 1993;268:2888–2892.
56. Lentz SR, Chen Y, Sadler JE. Sequences required for thrombomodulin cofactor activity within the fourth epidermal growth factor-like domain of human thrombomodulin. J Biol Chem 1993; 268:15312–15317.
57. Yang L, Rezaie AR. The fourth epidermal growth factor-like domain of thrombomodulin interacts with the basic exosite of protein C. J Biol Chem 2003;278: 10484–10490.
58. Wodak SJ, Janin J. Computer analysis of protein-protein interactions. J Mol Biol 1978;124:323–342.
59. van Dijk ADJ, de Vries SJ, Dominguez C, Chen H, Zhou H-X, Bonvin AMJJ. Data-driven docking: HADDOCK's adventures in CAPRI. Proteins 2005;60:232–238.
60. Carvalho AL, Dias FM, Prates JA, Nagy T, Gilbert HJ, Davies GJ, Ferreira LM, Romao MJ, Fontes CM. Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. Proc Natl Acad Sci USA 2003;100:13809–13814.
61. Terrak M, Kerff F, Langsetmo K, Tao T, Dominguez R. Structural basis of protein phosphatase 1 regulation. Nature 2004;429:780–784.
62. Graille M, Mora L, Buckingham RH, van Tilbeurgh H, de Zamaroczy M. Structural inhibition of the colicin D tRNase by the tRNA-mimicking immunity protein. EMBO J 2004;23:1474–1482.
63. Bressanelli S, Stiasny K, Allison SL, Stura EA, Duquerroy S, Lescar J, Heinz FX, Rey FA. Structure of a flavivirus envelope glycoprotein in its low-pH-induced membrane fusion conformation. EMBO J 2004;23:728–738.
64. Sansen S, de Ranter CJ, Gebruers K, Brijs K, Courtin CM, Delcour JA, Rabijns A. Structural basis for inhibition of *Aspergillus niger* xylanase by triticum aestivum xylanase inhibitor-I. J Biol Chem 2004;279:36022–36028.
65. Eghiaian F, Grosclaude J, Lesceu S, Debey P, Doublet B, Treguer E, Rezaei H, Knossow M. Insight into the PrPC→PrPSc conversion from the structures of antibody-bound ovine prion scrapie-susceptibility variants. Proc Natl Acad Sci USA 2004;101: 10254–10259.
66. Jacobs SA, Taverna SD, Zhang Y, Briggs SD, Li J, Eissenberg JC, Allis CD, Khorasanizadeh S. Specificity of the HP1 chromo domain for the methylated N-terminus of histone H3. EMBO J 2001;20:5232–5241.
67. Liu Q, Jin C, Liao X, Shen Z, Chen DJ, Chen Y. The binding interface between an E2 (UBC9) and a ubiquitin homologue (UBL1). J Biol Chem 1999;274:16979–16987.
68. Tatham MH, Kim S, Yu B, Jaffray E, Song J, Zheng Z, Rodriguez MS, Hay RT, Chen Y. Role of an N-terminal site of Ubc9 in SUMO-1, -2, and -3 binding and conjugation. Biochemistry 2003; 42:9959–9969.