PROTEINS: Structure, Function, and Genetics 44:336–343 (2001)

Prediction of Protein Interaction Sites From Sequence Profile and Residue Neighbor List

Huan-Xiang Zhou* **and Yibing Shan**
Department of Physics, Drexel University, Philadelphia, Pennsylvania

ABSTRACT **Protein–protein interaction sites are predicted from a neural network with sequence profiles of neighboring residues and solvent exposure as input. The network was trained on 615 pairs of nonhomologous complex-forming proteins. Tested on a different set of 129 pairs of nonhomologous complex-forming proteins, 70% of the 11,004 predicted interface residues are actually located in the interfaces. These 7732 correctly predicted residues account for 65% of the 11,805 residues making up the 129 interfaces. The main strength of the network predictor lies in the fact that neighbor lists and solvent exposure are relatively insensitive to structural changes accompanying complex formation. As such, it performs equally well with bound or unbound structures of the proteins. For a set of 35 test proteins, when the input was calculated from the bound and unbound structures, the correct fractions of the predicted interface residues were 69 and 70%, respectively. Proteins 2001;44:336–343.**

Key words: protein–protein interface; neural network; structural genomics; docking; homology modeling

INTRODUCTION

Protein–protein interactions play a central role in a number of biological processes such as immune response, enzyme catalysis, and signal transduction. Important goals of protein science are to understand the mechanisms of protein–protein interactions and to predict the interaction sites on the protein surfaces. Much interest has been focused on the docking problem, in which, given the unbound structures of two interacting proteins, one tries to locate the interaction sites on the two partners and then build a model for the protein–protein complex.[1–5] Most docking methods are based on the observation that protein–protein interfaces are composed of relatively large surfaces with geometric and perhaps electrostatic complementarity.[6,7] Progress in this direction has been greatly hampered by the conformational changes that usually accompany complex formation. Here we address the question of predicting interaction sites on one protein from its unbound structure without knowing the structure of its partner.

With cross-genome sequence comparisons, hundreds and thousands of putative protein–protein interaction pairs have been identified.[8–10] Experimentally, Uetz et al.[11] recently identified 957 interaction pairs in *Saccharomyces cerevisiae* with exhaustive two-hybrid screens. In the meantime, structural genomics and homology modeling efforts will likely generate structural models for most proteins in the next 10 years.[12,13] The next grand challenge will then be to find interaction sites on the proteins and build structural models for the protein complexes. The interface predictor developed here is ideally suited for that purpose.

It is known that there is a large presence of hydrophobic residues on interaction sites with respect to protein surfaces as a whole. It has even been suggested that the binding energy of two proteins derives from the burying of hydrophobic surface areas.[14] In general, Leu, Ile, Val, Phe, Tyr, and Met are overpopulated, whereas Lys, Asp, Glu, and other polar residues (with the exception of Arg) are underpopulated in interfaces.[15] In many interfaces, a hydrophobic core is surrounded by a ring of polar residues.[16] In an earlier attempt at interface prediction, Jones and Thornton[17,18] analyzed protein surfaces in terms of patches and showed that a score consisting of solvation potential, residue interface propensity, hydrophobicity planarity, protrusion, and accessible surface area was promising for predicting whether a surface patch overlapped with the interface.

The specific problem that this article addresses is as follows: Given the unbound structure of a protein and the fact that it does form a complex with an unknown protein, predict the residues of the first protein that will be located in the interface with the second protein. These interface residues define the sites of interactions with the second protein. Our predictor uses sequence profiles of neighboring residues and their solvent exposure to train a neural network. The rationale for grouping neighboring residues is that the interface is formed by one (or sometimes a few) spatially contiguous set of residues. This is analogous to considering adjacent residues along the peptide chain together in predicting secondary structures.[19] The solvent exposure is included to account for the fact that the residues eventually forming the interface are mostly exposed to the solvent prior to complex formation.

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

*Correspondence to: Huan-Xiang Zhou, Department of Physics, Drexel University, Philadelphia, PA 19104. E-mail: hxzhou@einstein.drexel.edu

Received 28 November 2000; Accepted 20 April 2001

The characteristics indicative of interaction sites can be captured by the position-specific sequence profiles from multiple-sequence alignment by PSI-BLAST.[20] For example, if a particular position along the sequence is mostly occupied by residues favored to be in protein interfaces, the chance of that position to be in the interface is high. A neural network can be trained to learn such simple and other more subtle tendencies. The network approach has been shown to be quite successful in predicting protein secondary structures.[21–23]

The network predictor for interface residues is found to have an accuracy of 70%. The prediction is not based on the protein sequence alone. It uses as input structural information, such as which residues are on the protein surface and for each surface residue which residues are its spatial (as opposed to sequential) neighbors. The input quantities chosen are relatively insensitive to the structural changes accompanying complex formation. Hence, the predictor has the strength that it performs equally well with either bound or unbound structures.

MATERIALS AND METHODS

Collection of Protein–Protein Complexes

All multiple-chain protein entries in the Protein Data Bank (PDB; June 2000 release) were examined to collect interfaces. For each PDB entry, distances between heavy atoms of any two chains were calculated. A residue is said to form an interfacial contact if the distance between any of its heavy atoms and any heavy atom from a partner chain is less than 5 Å. A pair of chains were retained when each had at least 20 residues that formed at least one interfacial contact with the other chain, with the provision that each chain could not appear in more than one retained pair. A total of 3704 pairs of chains were collected.

Each of the 3704 pairs of the sequences was aligned against all other sequences in the set by PSI-BLAST. Two chains were considered to have high homology if (1) over 90% of their sequences were matched and (2) the sequence identity over the matched region was greater than 40%. All chains with high homologies were collected in a cluster. The initial 3704 pairs of chains were then mapped to these clusters. Out of all the pairs with one chain mapped to cluster j and the second chain mapped to cluster k, one representative pair was chosen. The representative pairs form a nonredundant set of interacting protein chains. There are a total of 744 such pairs. Among these 564 are homodimers (or, more specifically, composed of chains sharing high homology), and 180 are heterodimers.

The 744 representative pairs of interacting proteins were divided into a training set of 615 pairs and a testing set of 129 pairs. The training set contained 63 heterodimers, whereas the test set contained 117 heterodimers. Many of the homodimers in the training set are probably formed only in the crystalline environment, and the proteins would be monomeric in solution. However, we included these in the training set to make it adequately large for the training purpose.

In the literature, the cutoff for high homology is sometimes set at 25% sequence identity. The main motivation for using the 40% cutoff here was again to collect enough representative chains so that the training set was adequately large. Then, there is the concern of whether the test set shares too much homology with the train set. Out of the 258 chains (i.e., 129 pairs) in the test set, only 56 could be aligned with those in the train set with over 90% of the sequences and with identities greater than 25%. We thus believe that the test set is sufficiently distinct from the training set.

Collection of Surface Residues

Interfaces are formed mostly by residues that are exposed to the solvent if the partner chain is removed. Therefore, we focused on those residues with accessible surfaces areas above certain thresholds. The accessible areas were calculated with the DSSP program.[24] In the calculation, only coordinates of the particular chain was used. That is, all other chains in the PDB file were stripped (otherwise, the surface areas of the residues that eventually form the interface with another chain would be incorrectly calculated).

The threshold for deciding whether a residue was a surface residue was set at 10% of the nominal maximum area for that type of residue. The nominal maximum areas were taken to be 106, 248, 157, 163, 135, 198, 194, 84, 184, 169, 164, 205, 188, 197, 136, 130, 142, 227, 222, and 142 $Å^2$ for Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, and Val, respectively. According to the aforementioned criterion, the 615 pairs of proteins allocated for training purpose have 225,139 surface residues. Only these surface residues were used for training the neural network.

For the training set, we classified a residue to be an interface residue if it formed at least three interfacial contacts with the partner chain. Of the 225,139 surface residues in the training set, 42,797 (19%) are interface residues. In addition, 3164 interface residues do not pass their surface area thresholds and thus were excluded for training the neural network. Such a price seems to be worth paying, as the surface area criterion allows us to detect interface residues among only 66% of the total 341,205 residues.

We used the criterion of at least three, as opposed to just one, interfacial contacts for designating a residue an interface site in order to reduce somewhat the chance of predicting an interface site. In this way, the prediction that a residue is an interface site can be more certain, and so the prediction accuracy can possibly be increased slightly. However, for the purpose of evaluating whether or not a prediction is correct, we do use the criterion of at least one interfacial contact for designating a residue an interface site (see the Results section).

For each surface residue, the distances with all other surface residues in the same chain were calculated and sorted in ascending order. The identities of the nearest neighbors were later used for the input to the neural network.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

(a)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
...																				
F78	0	-5	-5	-5	3	-5	-5	-5	-4	3	-1	-5	-2	7	-5	-4	-3	-2	0	1
L79	**-2**	**-4**	**-5**	**-6**	**-2**	**-4**	**-5**	**-6**	**-5**	**3**	**5**	**-4**	**2**	**-2**	**-5**	**-4**	**-3**	**-4**	**-3**	**2**
V80	-2	-4	-5	-6	3	-4	-5	-6	-5	3	4	-4	2	-2	-5	-4	-3	-4	-3	2
C81	0	-5	-5	-5	8	-4	-5	-3	-5	0	-1	-4	1	-3	-4	-3	-1	-5	-3	5
F82	-4	-5	-5	-5	-4	-4	-5	-5	-1	-2	-2	-5	-2	7	-6	-4	-4	0	8	-3
S83	0	-3	-1	5	-4	-2	-1	-2	-3	-4	-5	-2	-4	-5	-3	5	-2	-6	-5	-4
V84	-2	-5	-5	-5	-1	-4	-5	-6	-5	5	1	-4	0	-2	-5	-4	-1	-5	-3	5
V85	-1	-3	2	1	-3	-3	-3	-3	-4	1	-2	-3	-2	-4	-3	0	5	-5	-3	1
S86	-1	1	3	3	-1	0	0	-3	-2	-4	-4	0	-3	-5	-3	4	-1	-5	-3	-3
...																				

(b)

Fig. 1. (a) BLOSUM62 substitution matrix and (b) sequence profiles for a stretch of residues in a protein chain outputted by PSI-BLAST. The line for L79 is highlighted because it is used as an example for calculating the conservation score.

Sequence Profiles From PSI-BLAST

Sequence profiles were obtained with three rounds of PSI-BLAST searches. The database consists of 348,901 protein sequences from Swissprot. Both e and h were set to 10^{-3}. The substitution matrix was BLOSUM62.[25] The BLOSUM62 matrix and a sample output for the sequence profiles are shown in Figure 1.

Architecture of Neural Network

The architecture of the neural network predictor is shown in Figure 2. The sequence profile of a surface residue and its solvent-accessible area (scaled by the nominal maximum area) and the same quantities for the 19 spatially nearest surface residues (a total of 21×20 variables) make up the input. The 420 input nodes are fed

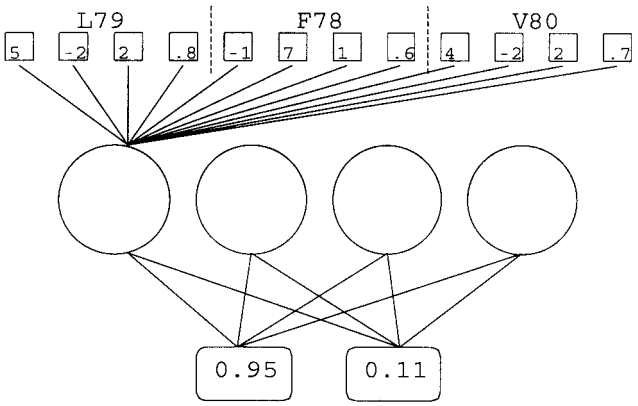

Fig. 2. Neural network predictor. The nodes of the input, hidden, and output layers are represented by squares, circles, and rounded rectangles, respectively. The residue under prediction is L79. For simplicity, only the 2 spatially nearest neighbors (not the 19 used in the actual prediction) are illustrated. Two dashed, vertical lines separate the input nodes for different residues. The sequence profile is also abbreviated to illustrate a hypothetical situation with just three types of residues (L, F, and V). The first 3 input nodes for each residue contain the scores of substitution (actually read from the sequence profile in Fig. 1 for L, F, and V). The last input node contains the solvent-accessible area scaled by the nominal maximum area (e.g., 0.8 for L79). The 12 input nodes should be connected to all 4 hidden nodes, but for clarity only connections to the first hidden node are shown. The values of the output nodes, 0.95 and 0.11, are indicated. The second neural network (not shown) has the same architecture, but has different types of input. The input from residue L79 is 0.95, 0.11, and 0.8. Corresponding values are also collected for F78 and V80. If the output values shown were those of the second neural network, L79 would be predicted an interface residue (since $0.95 > 0.11$).

to a hidden layer with 75 nodes, which in turn are fed to 2 output nodes.

In the training process, the targeted values for the two output nodes are (1,0) if the residue under prediction is an interface residue and (0,1) if that residue is not an interface residue. Training was carried out with a standard backpropagation procedure[26] on the 225,139 surface residues of the training set. For actual predictions, one compares the values of the two output nodes (x_1 and x_2, respectively). An interface site is predicted if $x_1 > x_2$, and a noninterface site is predicted otherwise.

We actually used two neural networks consecutively, as for predicting secondary structures.[21–23] The second neural network was included to improve prediction accuracy. We can explain the rationale by looking at two different scenarios. In both cases, residue i has been predicted by the first network to be in the interface. However, in the first case many of its spatial neighbors on the protein surface are also predicted by the first network to be in the interface, but in the second case none of the neighboring residues are predicted to be in the interface. Obviously, the chance that residue i is indeed in the interface will be much higher in the first case.

The second network has 60 input nodes, a hidden layer with 30 nodes, and 2 output nodes. The 60 input variables are the outputs from the first network and the accessible areas for a residue and its 19 nearest neighbors on the surface. It is the values of the 2 output nodes in the second network that one finally compares to predict whether a residue is an interface site.

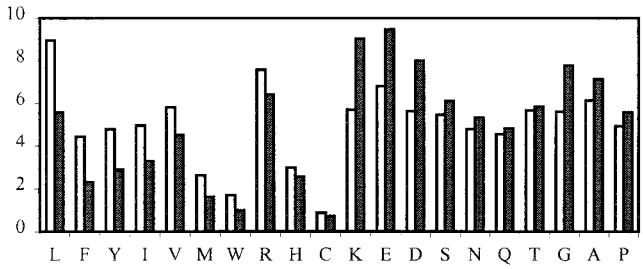

Fig. 3. Distributions of the 20 types of residues in the interface sites (open bars) and in the noninterface sites (shaded bars). For each type of residue, its percentage among all the residues that are designated as interface (or non-interface) sites is shown.

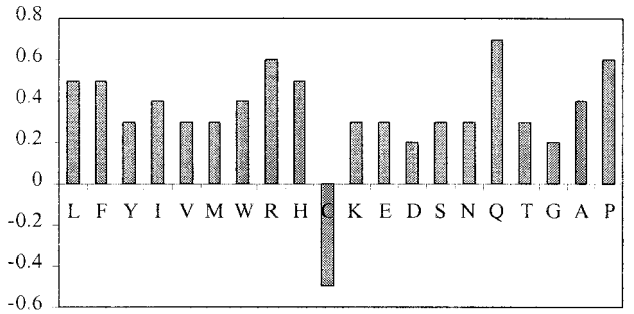

Fig. 4. Differences in the average diagonal elements of the sequence profiles between interface sites and noninterface sites.

RESULTS

Characteristics of Interface Residues

The distributions of the 20 types of residues among the 42,797 interface sites of the training set are shown in Figure 3 and compared with the distributions among the remaining 182,342 noninterface surface sites. It is immediately clear that nonpolar residues are favored in the interface, whereas charged and polar residues (except for Arg) are disfavored in the interface. The Leu fraction is higher by 3.4%, whereas the Lys fraction is lower by 3.3%. Gly also disfavors interface sites, with its fraction lower by 2.2%. These results are in agreement with the previous findings of Lijnzaad and Argos.[15]

Because interface residues are involved in protein–protein interactions, we expect them to mutate less frequently (i.e., be more conservative) than other surface sites. This is borne out by the PSI-BLAST sequence profiles. The diagonal element of the sequence profile at each residue position signifies the mutability of that residue: the higher that element, the less frequent its mutation. We averaged the diagonal elements for all Ala residues in interface positions and compared them against the corresponding average in noninterface surface positions. Figure 4 shows the differences for Ala and the other 19 types of residues. Except for Cys, the averages over the interface sites are all higher than those over the noninterface surface sites.

These characteristics indicate that strong signals are present in the sequence profiles. This gave us confidence that the neural network predictor would do well.

TABLE I. Prediction Results for 35 Unbound Structures

				Bound		Unbound	
Bound	Unbound	Description	RMSD (Å)	m	Accuracy	m	Accuracy
1azsB:A	1ab8A	C1A:C2A domains of adenylyl cyclase	1.5	52	79	45	69
1b6cA:B	1fkj	FKBP12:TGF-β receptor	0.5	2	100	1	100
1bqqT:M	1br9	TIMP-2:CDMT1-MMP	2.7	48	79	56	55
1bthQ:K	1bpi	BPTI:thrombin	1.3	19	89	17	94
1ceeA:B	1aje	CDC42:WASP	4.3	22	86	33	76
1cgiE:I	1chg	α-chymotrypsinogen:trypsin inhibitor	1.3	9	67	30	47
1cgiI:E	1hpt		1.8	1	100	6	100
1d0gB:T	1dg6A	APO2L:death receptor 5	2.6	51	25	39	33
1dfjI:E	2bnh	ribonuclease inhibitor:ribonuclease A	1.5	0	—	3	67
1dhkA:B	lose	porcine α-amylase:lectin-like inhibitor	1.4	5	100	7	71
1eaiB:D	1lvy	elastase:elatase inhibitor	0.7	10	30	5	80
1finC:D	1hcl	CDK2:cyclinA	4.1	14	100	12	83
1finD:C	1vin		0.4	46	59	10	80
1fosF:E	1junA	c-JUN:c-FOS	1.2	11	100	22	100
1frtB:A	1bmg	β-2 microglobulin:FC receptor	1.3	32	97	19	95
1hrtI:H	1hic	hirudin:α-thrombin	2.1	2	100	6	100
1ibrA:B	1byuA	ran:importin β	3.9	23	96	9	67
1ibrB:A	1qgrA		2.5	36	25	13	54
1itbA:B	2i1b	interleukin-1β:interleukin-1 receptor	1.0	8	100	10	100
1kigH:I	1hcgA	factor XA:anticoagulant peptide	0.8	9	44	10	50
1kigI:H	1tcp		2.3	0	—	3	100
1mahA:F	1maaA	acetylcholinesterase:fasciculin2	0.7	1	100	3	100
1mahF:A	1fsc		0.7	25	44	14	71
1nfdD:C	1bec	α-β T cell receptor heterodimer	2.9	24	88	22	95
1pytA:C	1pca	procarboxypeptidase A:proproteinase E	0.6	0	—	2	50
1pytC:A	1fonA		3.2	10	70	23	26
1qctA:B	1afcA	growth factor:growth factor receptor	0.8	5	100	1	100
1tbaB:A	1ytbA	TBP:TAFII230	1.8	3	100	3	100
1tbqH:R	1awhB	thrombin:rhodniin	1.0	1	100	4	25
1tmqA:B	1jae	worm α-amylase:ragi inhibitor	0.4	11	82	5	50
1tmqB:A	1bluA		1.1	29	69	20	80
1wq1R:G	1ctqA	RAS:RASGAP	0.6	12	100	12	100
1yagG:A	1d0nA	gelsolin:actin	2.0	11	55	0	—
4proB:D	2alp	α-lytic protease:PRO region	0.3	18	61	26	77
4proD:B	2proA		2.7	7	100	10	100

Interface Predictions on 129 Pairs of Test Proteins

The 129 pairs of test proteins have a total of 58,890 residues, of which 40,914 are on the surfaces. Among the surface residues, 11,805 have at least one interfacial contact. An interface residue prediction was considered correct when either the residue or one of its four nearest neighbors was one of the aforementioned 11,805. In general, for a residue at sequence position i, the first two of the nearest spatial neighbors are residues at sequence positions $i - 1$ and $i + 1$, and the next two are either residues at sequence positions $i - 2$ and $i + 2$ or residues from other parts of the sequence that are in direct contact with residue i.

A total of 11,004 interface residue predictions were made. Of these, 7732, or 70%, are correct according to the aforementioned evaluation criterion. The correctly predicted residues account for 65% of the 11,805 residues making up the 129 interfaces. In this evaluation, only interfaces with partners contained in the test set were considered. Often a protein can interact with more than one partner. The predicted interface residues, when located in the interfaces with these other partners, were counted as incorrect. Hence, the 70% accuracy is an underestimate (discussed later). If only the surface residues that are in direct contact with the partner protein are counted as correct predictions, the accuracy is 51% (5583 out of 11,004). This is to be compared with the interface fraction of the surface residues, 29% (11,805 out of 40,914).

We wondered what the correct fraction of predicted interface residues would be if we directly used the characteristics of interface residues rather than going through a neural network. In particular, we tested the following simple model: in each chain, 27% (11,004 out of 40,914) of the most conserved surface residues are predicted as interface sites. Residue conservation is measured by the sequence profile. For example, in the BLOSUM62 matrix, Leu has positive substitution scores of 2, 4, 2, and 1 with Ile, Leu, Met, and Val, respectively. These four scores are added to yield a baseline of 9. For residue L79 in Figure 1, the scores for substitution by these residues are 3, 5, 2, and 2, respectively (a total of 12). The difference of this total and the baseline, $12 - 9 = 3$, is called the conservation

score. We calculated the conservation scores for all the surface residues of a protein chain and sorted them in descending order. If that chain has N surface residues, the first $0.27N$ of them are predicted to be interface residues. With this procedure, 5331 of the 11,004 predicted residues are correctly predicted according to the evaluation criterion described previously. The correct fraction, 48%, is almost the same as a random prediction. When we randomly picked 27% of the surface residues of every chain to be interface residues, the correct fraction was 47% (5160 out of 11,004).

The neural network is far superior to the direct use of the characteristics of interface residues. Undoubtedly, one of the advantages of the neural network is the fact that spatially neighboring surface residues are grouped for interface site prediction. The preference of a particular residue for the interface may be weak, but when several neighboring surface residues all show a preference, the chance of an interface site becomes much greater. The grouping of adjacent residues along the peptide chain in predicting secondary structures is a key component of the Chou and Fasman algorithm.[19]

It appears that a major contribution to the high accuracy of the neural network predictor is that it has learned when to predict many interface residues and when to predict very few interface residues, depending on the characteristics of the proteins. We recorded the number of interface residues predicted for each chain by the neural network and used this as input for random prediction. In other words, if the neural network predicted m residues to be in the interface for a chain (with N surface residues), we randomly chose m surface residues (as opposed to $0.27N$ as done previously) to be interface residues. The accuracy of the random prediction now increased significantly to 56%.

Prediction With Unbound Structures

Ultimately, any interface prediction method must use only unbound structures. We searched for unbound forms for the 129 pairs of protein chains in the PDB and found 35 such proteins. These are listed in Table I. When data specific for these 35 chains were collected from the test results on the 129 pairs of proteins, the interface prediction accuracy was found to be 69%, nearly identical to what was found for the test set as a whole.

We calculated the neighbor lists and accessible areas from these unbound structures and fed the resulting input to the neural network previously trained on the 615 pairs of proteins. The overall prediction accuracy for the 35 unbound structures was 70%, even slightly better than that with the bound structures.

The total number (m) of predicted interface residues and the accuracy (%) for the individual proteins are listed in Table I. Although for 1pytC the prediction with the bound structure is better, for 1eaiB and 1ibrB, just the opposite is true. Bound and unbound structures can have root-mean-square deviations (RMSDs) as large as 4.3 Å. No correlation between RMSDs and differences in prediction accuracy can be detected. Next, we focus on predictions with the unbound structures.

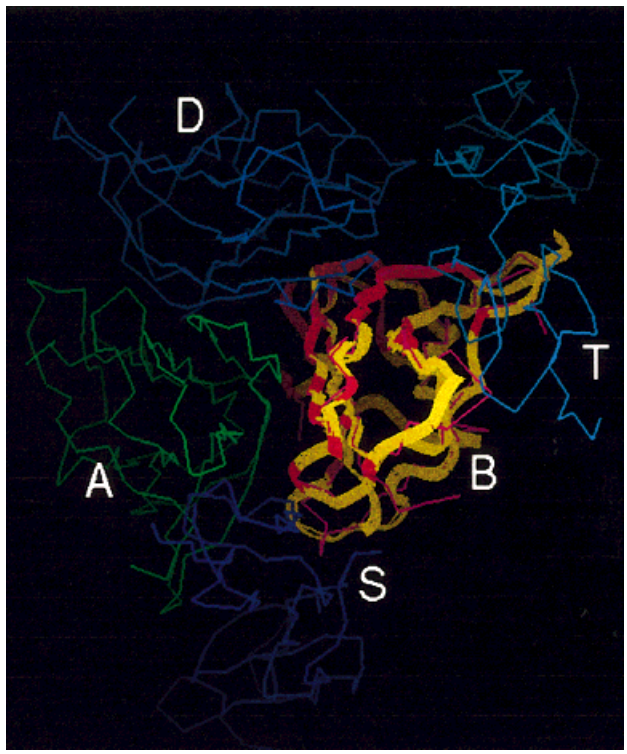

Fig. 5. Predicted interface residues (red ribbon) on 1d0gB (purple) with the unbound structure 1dg6A (yellow ribbon). The chosen partner chain is 1d0gT (cyan). Most of the predicted residues are located in the interfaces with the other chains (A, D, and S).

Locations of Predicted Interface Residues

An accuracy in terms of individual predicted residues of 70% is impressive. Examining the locations of the predicted residues relative to the interfaces of the protein complexes in graphic displays (InsightII, Molecular Simulations Inc.), we found that it is very conservative to include just four nearest neighbors in evaluating whether a prediction is correct. Many of the predicted residues not selected by this criterion as being correct are also seen in the interfaces. It appears that, when the accuracy is above 50% for a particular protein, the predicted interface residues are almost exclusively located around the interface.

In three cases (1d0gB, 1tbqH, and 1pytC), the accuracy is substantially below the 50% mark. After further examination of 1d0gB, we found that all the predicted interface residues not in the interface with the chosen partner (1d0gT) and thus not counted as correct are actually located in the interfaces with three other subunits (chains A, D, and S; see Fig. 5). The same situation happens in the case of 1tbqH, where the predicted interface residues not in the interface with the chosen partner (1tbqR) are found in the interface with chain L. 1tbqR represents the only case where the predicted interface is only half-right. The 23 predicted interface residues are clustered into two patches, one within the interface but the other on the opposite side of the surface. It is not clear whether this is an interaction site with a different partner. For 1yagG, no interface residues were predicted. Hence, interface predic-

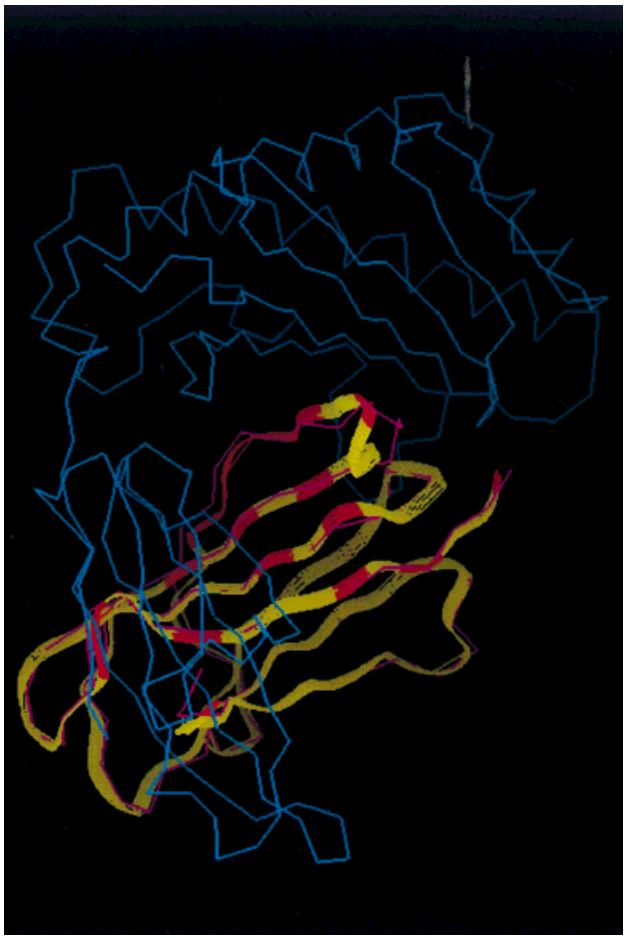

Fig. 6. Predicted interface residues (red ribbon) on 1frtB (purple) with the unbound structure 1bmg (yellow ribbon). The partner chain 1frtA is in cyan.

tions are completely successful in 33 of the 35 cases studied.

In most cases, the predicted residues are dispersed somewhat throughout the interface region and essentially cover the whole interface. This is illustrated by the prediction on 1frtB shown in Figure 6. The interaction sites in 1nfdD are located in two different domains; interface residues from both domains are correctly predicted. In other cases (e.g., 1b6cA, 1pytA, and 1qctA), although only a few interface residues are predicted, these residues are located around the top of an epitome or the rim of a valley, so the overall interface region can be clearly inferred.

Conformational changes with complex formation often occur most significantly around the interface. In 1finC (CDK2), the T-loop is flipped by about 150° after binding 1finD (cyclinA), and the T-loop becomes inserted into a crevice in 1finD.[27] With the unbound CDK2 structure (PDB entry 1hcl), two residues on the T-loop are correctly predicted as being in the interface (see Fig. 7). The conformational changes are so substantial that these residues appear to be away from the interface but will be in the interface if they are mapped to the bound structure. The PSTAIRE helix of CDK2 also shifts significantly and

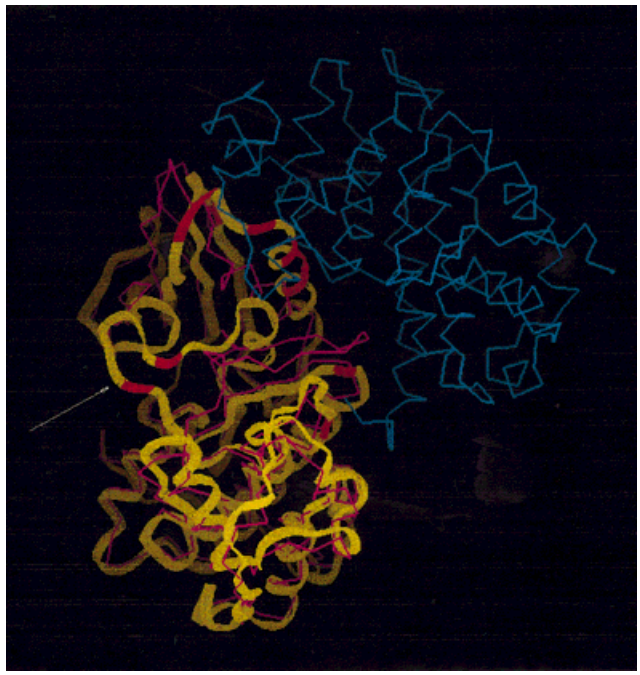

Fig. 7. Predicted interface residues (red ribbon) on CDK2 (purple) with the unbound structure 1hcl (yellow ribbon). The bound cyclinA is in cyan. The T-loop of CDK2 in the unbound form is indicated by an arrow. The PSTAIR helix is located above the bound T-loop, and the side facing cyclinA is correctly predicted to be in the interface.

comes into contact with cyclinA after complex formation. Even with the unbound structure, the side of the PSTAIRE helix facing cyclinA is completely predicted to be in the interface.

Substantial conformational changes also occur in the region of 1ceeA that comes into contact with 1ceeB. Here again the unbound structure (1aje) allowed for accurate interface prediction. All 33 predicted residues are either within the interface or in the immediate vicinity.

DISCUSSION

We have developed an accurate predictor for interface residues. The overall accuracy for all the predicted residues is 70% (or even higher when interfaces with multiple partners are considered). Graphic displays indicate that, for an individual protein, if the accuracy is above 50%, the predicted interface residues are indeed located in the interface almost exclusively.

The interface predictor uses sequence profiles of neighboring residues and their solvent exposure as input. These quantities are relatively insensitive to the structural changes usually accompanying complex formation. Hence, the predictor has the strength that it performs equally well with either bound or unbound structures. This strength is demonstrated on 35 proteins.

The predicted interface residues can be directly used to guide experimental studies. For example, they may direct experimental efforts toward a particular region on a protein surface in studying its interactions with another protein. They may also be used to help solve the docking

problem by vastly reducing the amount of configurational space needed to be searched to build a structural model for the protein complex.

This study, similar in spirit to the previous work of Jones and Thornton,[17,18] extends that work in a number of significant ways. First, we characterized the protein surface and made interface predictions at the level of individual residues as opposed to surface patches. Second, we used sequence profiles rather than a single sequence to capture the characteristics of interface residues. Third, our prediction is based on training a neural network on a set of 615 protein pairs and is tested on a different set of 129 protein pairs, whereas in the method of Jones and Thornton, empirical rules were learned from 59 protein structures and tested on the same proteins. Unfortunately, these differences make a direct comparison of the two methods impossible.

Because of the scarcity of dimeric protein structures in the PDB, we have used dimer interfaces formed only in the crystalline environment (mainly for the purpose of training). However, we selected only those interfaces that have extensive interfacial contacts (specifically, involving at least 20 residues from each side). These large interfaces are expected to be stabilized by the same types of interactions that stabilize dimer interfaces in solution.

The interface predictor developed here does not use information from the partner protein. Correlated mutations between two domains of a multiple-domain protein have been shown to embody important information about interaction sites.[28] For two isolated chains, the only situation where correlated mutations can be used is when homologues of the fusion protein of the two chains exist extensively in the sequence database.

In general, proteins function by interacting with other proteins. Many such interaction pairs have now been identified by computational and experimental means.[8–11] Many of the structures of the partner proteins have been predicted by homology modeling e.g., Shan et al.[23]). The method developed here can then be used on these modeled structures (equivalent to the unbound structures used in this study). Our interface predictor thus appears to be ideally suited for taking on the next grand challenge of finding interaction sites on the proteins and building structural models for the protein complexes. The server for this predictor, PPISP, is located at http://cmbph1.physics.drexel.edu/cgi-bin/PPISP.cgi.

REFERENCES

1. Shoichet BK, Kuntz ID. Protein docking and complementarity. J Mol Biol 1991;221:327–346.
2. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci U S A 1992;89:2195–2199.
3. Norel R, Lin SL, Wolfson HJ, Nussinov R. Molecular surface complementarity at protein–protein interfaces: the critical role played by surface normals at well placed, sparse, points in docking. J Mol Biol 1995;252:263–273.
4. Gabb HA, Jackson RM, Sternberg MJ. Modeling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 1997;272:106–120.
5. Palma PN, Krippahl L, Wampler JE, Moura JJG. BiGGER: a new (soft) docking algorithm for predicting protein interactions. Proteins 2000;39:372–384.
6. Janin J. Principles of protein–protein recognition from structure to thermodynamics. Biochimie 1995;77:497–505.
7. Jones S, Thornton JM. The principles of protein–protein interactions. Proc Natl Acad Sci U S A 1996;93:13–20.
8. Marcotte EM, Pellegrini M, Ng H-L, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein–protein interactions from genome sequences. Science 1999;285:751–753.
9. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. Nature 1999;402:83–86.
10. Enright AJ, Ilipoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature 1999;402:86–90.
11. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Queshi-Emilli 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;403:623–627.
12. Montelione GT, Anderson S. Structural genomics: keystone for a human proteome project. Nat Struct Biol 1999;6:11–12.
13. Zhou H-X, Wang G. Projecting the years of structure determination needed for complete annotation of 32 genomes. http://cmbph1.physics.drexel.edu/projecting.html.
14. Chothia C, Janin J. Principles of protein–protein recognition. Nature 1975;256:705–708.
15. Lijnzaad P, Argos P. Hydrophobic patches on protein subunit interfaces: characteristics and prediction. Proteins 1997;28:333–343.
16. Larson TA, Olson A, Doodsell D. Morphology of protein–protein interfaces. Structure 1998;6:421–427.
17. Jones S, Thornton JM. Analysis of protein–protein interaction sites using surface patches. J Mol Biol 1997;272:121–132.
18. Jones S, Thornton JM. Prediction of protein–protein interaction sites using patch analysis. J Mol Biol 1997;272:133–143.
19. Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry 1974;13:222–245.
20. 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.
21. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
22. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
23. 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.
24. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
25. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 1992;89:10915–10919.
26. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986;323:533–536.
27. Jeffrey PD, Russo AA, Polyak K, Gibbs E, Hurwitz J, Massague J, Pavletich JP. Mechanism of CDK activation revealed by the structure of a cyclin A-CDK2 complex. Nature 1995;376:313–320.
28. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein–protein interaction. J Mol Biol 1997;271:511–523.