Faculty Publications in the Biological Sciences

Papers in the Biological Sciences

2006

Protein Binding Site Prediction Using an Empirical Scoring Function

Shide Liang
SUNY Buffalo

Chi Zhang
University of Nebraska - Lincoln, zhang.chi@unl.edu

Song Liu
SUNY Buffalo

Yaoqi Zhou
SUNY Buffalo

Protein binding site prediction using an empirical scoring function

Shide Liang, Chi Zhang, Song Liu and Yaoqi Zhou*

Howard Hughes Medical Institute Center for Single Molecule Biophysics, Department of Physiology and Biophysics, State University of New York at Buffalo, 124 Sherman Hall, Buffalo, NY 14214, USA

ABSTRACT

Most biological processes are mediated by interactions between proteins and their interacting partners including proteins, nucleic acids and small molecules. This work establishes a method called PINUP for binding site prediction of monomeric proteins. With only two weight parameters to optimize, PINUP produces not only 42.2% coverage of actual interfaces (percentage of correctly predicted interface residues in actual interface residues) but also 44.5% accuracy in predicted interfaces (percentage of correctly predicted interface residues in the predicted interface residues) in a cross validation using a 57-protein dataset. By comparison, the expected accuracy via random prediction (percentage of actual interface residues in surface residues) is only 15%. The binding sites of the 57-protein set are found to be easier to predict than that of an independent test set of 68 proteins. The average coverage and accuracy for this independent test set are 30.5 and 29.4%, respectively. The significant gain of PINUP over expected random prediction is attributed to (i) effective residue-energy score and accessible-surface-area-dependent interface-propensity, (ii) isolation of functional constraints contained in the conservation score from the structural constraints through the combination of residue-energy score (for structural constraints) and conservation score and (iii) a consensus region built on top-ranked initial patches.

INTRODUCTION

Biological cells function through a network of interacting proteins and other molecules. It has been estimated that the average number of interacting partners per protein is three to ten (1). In order for a protein to interact dynamically with multiple partners, the complexes of interacting proteins are often not obligatory but necessary transient with relatively weak binding affinity. Such a weak binding affinity, however, makes it difficult to solve the structures of transient complexes experimentally. As a result, there is a growing gap between the number of known interactions and the number of their 3-dimensional structures that are available. However, the 3D structures of protein complexes are pivotal for a full understanding of the mechanism of interactions because they provide specific interaction details at the atomic level. Such details are important for rational design of drug molecules to modulate protein interactions.

One way to solve this problem is molecular docking (2,3). In molecular docking, transient complex structures are predicted by docking one monomeric structure (typically the smaller one) onto the other. It consists of two steps: conformational sampling that generates multiple candidate complex structures and scoring that recognizes the near-native complex structures from the candidate complex structures. Here, we define a complex structure as transient if there exist corresponding monomeric structures. The accuracy of protein–protein docking (4–8) can be improved significantly if their binding region is known. This is because identification of binding regions dramatically reduces the conformational space of docking. Several recent studies attempted to predict possible protein–protein binding sites (interface residues) from known unbound monomer structures (9,10).

To predict interface residues, one needs to know what distinguishes an interface region from the rest of the protein surface. It was discovered that the interfaces of obligate complexes are outstandingly hydrophobic (11). The interfaces of some transient complexes were also found to be with clusters of hydrophobic residues (12). Moreover, they are rich in aromatic residues and arginine but depleted in other charged residues (13). However, hydrophobicity at the interfaces of transient complexes is not as distinguishable from the remainder of the surface as hydrophobicity at the interfaces of the obligate complexes (13,14). As a result, it is difficult to make an accurate prediction of the interfaces of transient complexes using a single parameter of residue interface propensity. Moreover, different interface properties of obligate and transient complexes make it necessary to treat them separately in prediction.

Evolution conservation of residues is another widely-used property to identify protein–protein interfaces (15–19).

*To whom correspondence should be addressed. Tel: +1 716 829 2985; Fax: +1 716 829 2344; Email: yqzhou@buffalo.edu

Interface residues, especially those hot spot residues (20) were found to be more conserved than other surface residues. However, residue conservation is rarely sufficient for a complete and accurate prediction of protein interface (21–23). Moreover, transient interfaces evolve faster than obligate ones (24). A more sensitive evolutionary tracing (ET) method, which delineates invariant residues responsible for subgroup accuracy, has been developed to uncover functionally important residues in proteins (6,25,26).

Several authors studied residue-energy distributions on the protein surface to identify functional sites. Hertzberg and Moult found that steric strains in the polypeptide backbone are located overwhelmingly in regions concerned with function (27). Elcock (28) predicted functionally important residues based on the assumption that they have a high electrostatic energy, as calculated with continuum methods. Ota *et al.* (29) combined stability profile and sequence conservation to predict catalytic residues in enzymes. Chelliah *et al.* (30) predicted binding sites by distinguishing restraints that arise from the structure of the protein and those from intermolecular interactions. Cheng *et al.* (31) used all-atom computational protein design methodology to compute the free energy difference between the naturally occurring amino acids and the lowest free-energy amino acids. Functional sites were found to have residues with sub-optimal free energies.

Other interface-distinguishing features have also been exploited. For example, Jones and Thornton (32) found that protein interfaces are among the most planar and most accessible patches. In structural terms, a binding site has a preference for β-sheets and relatively long non-structured chains, but not for α-helices (10). It was also shown that the binding site of an unbound monomer is surrounded by more bound water molecules and has a lower temperature B-factor than the rest of the surface (10). This result is consistent with the finding that an interfacial surface region is less flexible than the rest of the protein surface (33). More rigid interface region in unbound structures suggests that interface residues are 'prepared' for the loss of side-chain conformational entropy upon binding. In another study, Fernandez and Scheraga (34) found that most backbone hydrogen bonds in the majority of soluble proteins are thoroughly wrapped intramolecularly by non-polar groups except for a few that are likely around a binding site.

Because none of the above-mentioned properties are able to make an unambiguous identification of interface regions or patches (32), a combination of some of them [via either a linear combination (9,10) or machine learning (35–37)] was found to be effective for improving the accuracy of binding-site prediction. However, the overall accuracy remains low compared to expected random values.

This study is spurred by the finding that residues at the interfaces of protein–protein complexes have higher side-chain energies (i.e. less stable) than the other surface residues (38) based on a score function originally developed for protein design (39). The high energies of the observed interfaces are independent of the amino acid composition in the interface regions. Moreover, residues with high-energy scores are unstable and likely evolve fast. This suggests that the energy score, a suitable indicator of interface residues, is 'orthogonal' to properties describing interface propensity

and residue conservation. In other words, combination of the three terms might improve the accuracy of interface prediction.

In this paper, we test this hypothesis by developing a method, called Protein Interface residUe Prediction (PINUP). PINUP predicts interface residues using an empirical score function made of a linear combination of the energy score, interface propensity and residue conservation score. To our knowledge, none of the previously published work has combined these terms together to predict protein binding sites. We show that PINUP provides a significant improvement in the accuracy of binding site prediction over any single or pairwise combinations of the three terms. The PINUP server is freely available on http://theory.med.buffalo.edu/PINUP.

MATERIALS AND METHODS

The scoring function for interface-residue identification

The scoring function for a surface residue i, $E(i)$, is made of three terms: side-chain energy score, $E_{\text{sidechain}}(i)$, residue-conservation score, $E_{\text{consv}}(i)$ and residue interface propensity $E_{\text{propensity}}(i)$. That is,

$$E(i) = E_{\text{sidechain}}(i) + w_c E_{\text{consv}}(i) + w_p E_{\text{propensity}}(i), \quad \textbf{1}$$

where w_c and w_p are to-be-determined weight factors for the conservation and propensity scores, respectively.

1. Side chain energy score. We use a scoring function that was originally developed for protein design (39) to calculate the energy of a rotamer (40), the representative conformation of the amino acid, placed on its backbone position. The score for a given rotamer R of a residue i, $E_{\text{sidechain}}(R_i)$, is a linear combination of multiple energetic terms:

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

$$\textbf{2}$$

where S_{contact}, V_{overlap}, E_{hbond}, E_{elec}, ΔS_{pho} and ΔS_{phi} are atom-contact surface area, overlap volume, hydrogen bonding energy, electrostatic interaction energy, buried hydrophobic solvent accessible surface and buried hydrophilic solvent accessible surface between the rotamer of residue i and the rest of the protein, respectively; F_{phi} is the fraction of the buried surface of non-hydrogen-bonded hydrophilic atoms; $\Delta (F_{\text{phi}})^{30}$ is the difference between the rotamer positioned in the protein environment and the isolated form; $V_{\text{exclusion}}$ is the normalized solvent exclusion volume around charged atoms; f_1 is the observed frequency of the rotamer and f_2 is the observed frequency of the amino acid residues in a given backbone conformation; N_{ssbond} is the flag of disulfide bridge(1 or 0); ΔG_{ref} is the difference between the free energy of the rotamer in solvent and denatured protein. The weights of these energy items together with the reference values were optimized so that the native residue was predicted energetically favorable at each position of the training proteins (39).

The energy score of a particular amino acid i is calculated as:

$$E_{\text{sidechain}}(i) = -f \ln \left\{ \sum_{R} \exp\left[-E_{\text{sidechain}}(R_i) \right] \right\}, \qquad \textbf{3}$$

where the summation is over all the rotamers available for a given residue type and the constant prefactor f is $1/2.41$. This constant factor is from the slope of the regression line between the calculated and experimentally-measured unfolding $\Delta\Delta G$ of a set of point mutation data (39). Thus, the energy unit is $\text{kcal} \cdot \text{mol}^{-1}$, the same as the experimental data.

2. Residue conservation score. Residue conservation is measured by the self-substitution score from the sequence profile. Sequence profiles are obtained by three iterations of PSI-BLAST searches against non-redundant (NR) database posted on Dec 5, 2005 (ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr/gz) with the BLOSUM62 (41) substitution matrix. The conservation score at the position i is defined as

$$E_{\text{consv}}(i) = \begin{cases} M_{ir} - B_{rr}, & \text{if } M_{ir} - B_{rr} > 0, \\ 0, & \text{if } M_{ir} - B_{rr} \leq 0, \end{cases} \qquad \textbf{4}$$

where M_{ir} is the self-substitution score in the position-specific substitution matrix generated from PSI-BLAST for the residue type r at sequence position i and B_{rr} is the diagonal element of BLOSUM62 for residue type r. A slightly different criterion based on positive substitution scores has been used in Ref. (15). Here, we use the difference in self-substitution scores for simplicity.

3. Residue interface propensity. We define a residue-interface propensity, $E_{\text{propensity}}(i)$, that depends on its accessible surface area.

$$E_{\text{propensity}}(i) = \left(\ln \frac{P_r^{\text{interface}}}{P_r^{\text{surface}}} \right) \cdot \frac{S_r}{S_r^{\text{ave}}}, \qquad \textbf{5}$$

where $P_r^{\text{interface}}$ and P_r^{surface} are the contribution of residue type r to the interface area and to the protein surface area, respectively, and S_r and S_r^{ave} are the relative accessible surface area of residue r at the sequence position i and the average relative accessible surface area of surface residues of type r, respectively. The C_α atom of Gly is considered as a side chain atom for convenience. $P_r^{\text{interface}}$ and P_r^{surface} are obtained from statistical analysis of 75 protein–protein complexes (13). S_r^{ave} for 20 amino acid residues are obtained from statistical analysis of eight unbound monomers of four protein–protein complexes belonging to different families: barnase-barstar, hen egg white lysozyme-antibody Fab D44.1, acetylcholinesterase-fasciculin and chymotrypsin-OMTKY3. We found that eight unbound monomers yield enough statistics for calculation of S_r^{ave}. Details can be found in the Results section. The values of $\left(\ln P_r^{\text{interface}}/P_r^{\text{surface}} \right)/S_r^{\text{ave}}$ are listed in Table 1.

PINUP algorithm for predicting the interface residues

The PINUP algorithm is as follows:

(i) Identification of surface residues. As in a previous study (38), surface residues are defined as those side chains with a relative accessibility of >6% (probe radius = 1.2 Å).

Table 1. The values of $(\ln P_r^{\text{interface}}/P_r^{\text{surface}})/S_r^{\text{ave}}$ for 20 amino acid residues

Amino acid	$(\ln P_r^{\text{interface}}/P_r^{\text{surface}})/S_r^{\text{ave}}$	Amino acid	$(\ln P_r^{\text{interface}}/P_r^{\text{surface}})/S_r^{\text{ave}}$
Ala	−0.925	Leu	1.07
Arg	0.291	Lys	−0.991
Asn	−0.248	Met	2.22
Asp	−0.571	Phe	3.00
Cys	2.78	Pro	−0.553
Gln	−0.685	Trp	4.39
Glu	−0.881	Val	0.278
Gly	0.042	Ser	−0.749
His	1.56	Thr	−0.730
Ile	2.46	Tyr	3.76

(ii) Identification of candidate binding surface patches. A surface patch is defined as a central surface residue and 19 nearest neighbors as in a previous study (38). The score of a patch is given by the average value of the scores for all 20 residues by using the above-described scoring function. All of the surface residues are sampled. Solvent vector constraints (32) are applied in order to avoid patches sampling at different sides of a protein surface. Top 5% scored patches are selected. If the number of surface residues for a protein is less than 100, five top-scored patches are selected, instead.

(iii) Locating candidate interface residues. Typically, the above selected patches overlap with each other. That is, one residue can appear in multiple patches. We rank residues based on the number of top-scored patches to which they belong (the appearance rate in top-scored patches). Top-ranked 15 residues are designated as candidate interface residues. For large proteins with more than 150 surface residues, we retain up to 10% of total surface residues. If the last candidate residue (e.g. the 15th residue for proteins with less than 150 residues) has the same appearance rate in the top-scored patches as several other non-candidate residues, all of them are included in the candidate interface residues.

(iv) Prediction of a continuous binding interface. The final predicted interface is defined as the largest continuous patch made of the 'interacting' candidate interface residues. Two residues are considered interacting if the distance between any two respective side chain atoms is <1 Å plus the sum of the van der Waals radius of the two atoms. If a surface residue is surrounded by the predicted interface residues and it does not interact with other surface residues, the residue will be included as interface residues. The van der Waals radii for all atom types are from the CHARMM21 parameter set (42).

There are several parameters, such as the definition of surface residues [Step (1)] and the size of surface patches [Step (2)] in this PINUP algorithm. Effects from varying these parameters are discussed in the Results section.

Protein datasets

We use a set of 57 non-homologous proteins collected by Neuvirth *et al.* (10) for training and cross validation. In this

set, antibodies and antigens are not included since their specific binding mode is optimized through rapid somatic cell mutations instead of evolution over many years. Our algorithm relies on conservation information and, thus, is not suitable for predicting antigen–antibody interfaces. The structures of the unbound monomers and complexes are obtained from PDB (43). The program REDUCE (44) is used to add hydrogen atoms to all proteins. Non-polar hydrogen atoms and all water molecules are deleted. The binding sites are predicted with unbound structures. The complex structures are used to define the experimental interface residues for the unbound monomers. A surface residue is considered as interface residues if its accessible surface area is decreased by more than 1 Å^2 upon complexation.

To further test PINUP, we use the protein–protein docking benchmark 2.0 established by Chen *et al*. (45). This benchmark contains 62 protein complexes (excluding antigen–antibody), in which 68 unbound proteins can be considered as an independent test set because they share <35% sequence identity with any protein in the 57-protein dataset described above. This 68-protein set contains 42, 18 and 8 proteins with minor, medium and large-scale conformational change upon complexation, respectively.

There is a significant homologous relation between the 75 proteins used for deriving interface propensity and the 57 proteins used for cross validation. We test the dependence of prediction accuracy on the dataset used for deriving interface propensity and find that the dependence is essentially negligible. Details can be found in the Results section.

Assessment of prediction accuracy

Prediction accuracy is assessed by the coverage of the actual interface by the predicted interface, which is the fraction of correctly predicted interface residues in the total number of observed interface residues, and the accuracy of the predicted interface, which is the fraction of correctly predicted interface residues in the total number of predicted interface residues. The expected accuracy from random prediction is the fraction of observed interface residues in the total number of surface residues.

Optimizing the weights

We use a simple grid method to optimize the weights of w_c and w_p. An initial scanning suggests the optimal values located at $0 < w_c < 2$ and $1 < w_p < 10$. The final weights are obtained by a simple grid search within $0 < w_c < 2$ with a step of 0.2 and $1 < w_p < 10$ with a step of 1. The parameters are optimized for the highest accuracy.

RESULTS

Training and cross validation

Side chain energy scores, residue conservation scores and interface propensity are utilized to predict protein–protein interface for the unbound protein structure. The weights of the three items are balanced to achieve the highest accuracy rate by a grid-based search method. Leave-one-out cross validation is performed on the 57 proteins. Briefly, one protein is

removed from the training set. The weights, optimized with the other 56 proteins, are used to predict the binding site of the removed protein. This procedure is repeated 57 times until all the proteins have been left out. In 55 out of the 57 cases, the optimized weights are 1, 1.6 and 6 for energy score, conservation score and interface propensity, respectively. Compared with training on the whole dataset, the average prediction accuracy is only decreased by 0.7% in the leave-one-out procedure. This indicates that the weights obtained are robust for other proteins.

The prediction accuracy for the 57 proteins from the leave-one-out testing is shown in Table 2. The average values of coverage and accuracy are 42 and 45%, respectively. The average size of predicted interfaces is 21 residues, similar to the average size of observed interfaces (22 residues). It is of interest to note that both coverage and accuracy for interface prediction of inhibitors are significantly higher than those of either enzymes or other proteins.

There are six proteins predicted with more than one interface region. PINUP only accepts the largest region. Nonetheless, both the largest and the second largest regions are overlapped with the observed interface in three of the six proteins. If both regions were considered as the predicted interface, the coverage would be higher.

There are six proteins for which none of the interface residues are correctly predicted. Among them, the unbound crystal structure of nitric oxide synthase oxygenase domain (1nos) has only one interface residue with coordinates. It is the second-ranked interface in PINUP having an overlap with the single interface residue. All other interface residues identified in the bound structure have no structure (invisible in crystal structure) in the unbound structure. That is, these interface residues are not possible to be predicted because they are not part of surface residues in the unbound structure. For the horse plasma gelsolin (1d0n), small number of actual interface residues in a large surface may have led to the failure of PINUP to locate its interface residues. This protein with 589 surface residues has a relatively small interface of 24 residues. The failure is also likely due to the difficulty to generate the surface patches with same shape and size as the observed interface for a given unbound structure. A closer examination indicates that the observed interface of 1d0n ranks at the top 1 in all generated surface patches of the same size. Another cause for completely failed predictions is the existence of other possible binding sites. The predicted protein binding site of aspartate transcarbamoylase (1ekx) is overlapped with N-phosphonacetyl-L-aspartic acid (PAL) ligand binding site, rather than the observed protein–protein interface. The three remaining proteins with failed interface predictions are domains rather than the whole protein. The interface for domain–domain interactions may have prevented correct predictions. For example, a domain–domain interaction site of Enteropathogenic *Escherichia coli* intimin C-terminal domain (1f00), is predicted as the protein–protein binding interface.

One interesting question is what happens if bound structures are used for interface prediction. We find that when the bound structure of 1nos is used, the prediction coverage and accuracy are 91 and 50%, respectively. However, in general, the prediction accuracy and coverage by using bound structures is slightly lower than by using the unbound

Table 2. Leave-one-out cross validation for 57 unbound protein structures

Unbound protein[a]	PDB code	Complex code	interface residues	Surface residues	Coverage (%)	Accuracy (%)
Barstar	1a19A	1brsA	17	70	82.4	73.7
Barnase	1a2pA	1brsD	18	91	72.2	61.9
Tumor suppressor p16ink4a	1a5e-	1bi7B	31	125	22.6	38.9
Acetylcholinesterase	1acl-	1fssA	25	355	56.0	35.9
Plastocyanin	1ag6-	2pcfA	23	78	39.1	60.0
cdc42hs	1aje-	1am4D	18	160	33.3	60.0
Rhogdi	1ajw-	1cc0E	13	127	69.2	69.2
Fkbp-rapamycin-binding domain	1aueA	1fapB	8	78	37.5	15.8
Trypsin inhibitor	1avu-	1avwB	15	138	93.3	66.7
Human procarboxypeptidase a2	1aye-	1dtdA	24	304	54.2	44.8
Hydrolase angiogenin	1b1eA	1a4yB	34	101	47.1	88.9
Bifunctional trypsin/alpha-amylase inhibitor (rbi)	1bip-	1tmqB	27	119	37.0	58.8
Cytochrome *f*	1ctm-	2pcfB	25	201	36.0	36.0
Granulocyte colony stimulating factor	1cto-	1cd9B	6	103	100.0	35.3
Receptor chey mutant	1cye-	1eayA	13	97	7.7	6.7
Calcium-free equine plasma gelsolin	1d0nA	1c0fS	24	589	0.0	0.0
Hydrolase inhibitor	1d2bA	1ueaB	22	109	59.1	81.2
Transferase	1ekxA	1d09A	21	232	0.0	0.0
Bovine chymotrypsinogen a	1ex3A	1cgiE	28	188	46.4	68.4
Neuronal t-snare syntaxin-1a	1ez3A	1dn1B	18	116	44.4	47.1
N-terminal domain of enzyme I from *E.coli*	1eza-	3ezaA	24	249	0.0	0.0
rgs4	1eztA	1agrE	21	115	19.0	23.5
Enteropathogenic *E.coli* intimin C-terminal domain	1f00I	1f02I	16	215	0.0	0.0
Coxsackie virus and adenovirus receptor	1f5wA	1kacB	18	98	22.2	23.5
Fk506 binding protein	1fkl-	1b6cA	20	88	45.0	69.2
Uracil-DNA glycosylase	1flzA	1euiA	26	172	69.2	78.3
Neuronal sec1	1fvhA	1dn1	41	423	39.0	55.2
Hydrolase	1g4kA	1ueaA	31	133	22.6	41.2
Radixin ferm domain	1gc7A	1ef1A	59	248	11.9	31.8
Granulocyte colony stimulating factor (rhg-csf)	1gnc-	1cd9A	18	174	16.7	17.6
N-terminal region of p67phox	1hh8A	1e96B	14	150	78.6	44.0
Lipase (EC 3.1.1.3)	1hplA	1ethA	21	325	9.5	8.0
p53 core DNA-binding domain	1hu8A	1ycsA	16	155	68.8	29.7
Interleukin-1 beta	1iob-	1itbA	42	133	19.0	47.1
Actin	1j6zA	1c0fA	30	281	30.0	33.3
α-Amylase	1jae-	1tmqA	33	316	45.5	78.9
(EC 3.5.1.28) mutant	1lba-	1aroL	16	112	37.5	40.0
Knob domain from adenovirus serotype 12	1nobA	1kacA	18	139	0.0	0.0
Nitric oxide synthase oxygenase domain	1nos-	1nocA	1	251	0.0	0.0
Porcine pancreatic procolipase b	1pco-	1ethB	18	85	22.2	18.2
Profilin	1pne-	1hluP	25	107	60.0	93.8
Phosphotransferase (hpr)	1poh-	1ggrB	16	69	75.0	66.7
Papain (EC 4.3.22.2)	1ppp-	1stfE	28	160	32.1	50.0
Streptokinase domain b	1qqrA	1bmlC	10	118	70.0	38.9
Rhogap	1rgp-	1am4A	17	155	41.2	35.0
Selenosubtilisin	1selA	1cseE	23	177	47.8	61.1
Cyclin a	1vin-	1finB	29	194	51.7	50.0
P120gap	1wer-	1wq1G	34	251	35.3	66.7
α-Lactamase tem1	1×pb-	1jtgA	33	186	54.5	90.0
Ribonuclease inhibitor	2bnh-	1a4yA	39	370	35.9	36.8
Cyclophilin a	2cpl-	1ak4A	19	120	63.2	75.0
Glucose-specific phosphocarrier	2f3gA	1ggrA	20	104	60.0	63.2
Negative factor (fprotein)	2nef-	1avzB	14	128	42.9	33.3
RalGEF-rbd streptomyces	2rgf-	1lfdA	15	79	33.3	26.3
Subtilisin inhibitor	3ssi-	2sicI	13	92	100.0	68.4
Cytochrome c peroxidase (EC 1.11.1.5) mutant	6ccp-	2pcbA	13	226	69.2	29.0
BLIP	Bound	1jtgB	30	136	40.0	63.2
Mean			22	174	42.2	44.5
Enzyme (25)			21	174	43	47.3
Inhibitor (7)			21	149	68.6	65.6
Others (25)			23	181	34.1	35.8

[a]The same dataset has been used by Neuvirth *et al.* (10) except BLIP. The unbound structure of BLIP was not available in PDB and we used the bound structure instead.

structures. One contributing factor is that the crystal structures of transient complexes (i.e. bound structures) have a lower resolution than the unbound monomers. Moreover, the coordinates of some surface residues (other than the interface residues) are often missing. This leads to misidentification of the exposed core region as surface residues and thus, reduction of the overall accuracy of interface prediction by including the exposed core region in the prediction.

Table 3. Effect of individual and combined scores

No.	Energy score[a]	Conservation score[a]	Interface propensity[a]	Coverage (%)	Accuracy (%)	Interface rank[b]	Difference[c]
1	1			19.2	18.7	2.7	0.46 (kcal·mol^{-1})
2		1		22.6	24.7	4.6	0.12
3			1	37.1	35.2	2.8	0.20
4	1	1.2		25.9	27.0	2.3	0.61
5	1		7	38.7	39.6	2.0	1.88
6		1	3	37.7	39.5	2.8	0.73
7	1	1.6	6	43.1	45.2	2.0	1.87

[a]The weights of the combined scores are optimized to achieve the highest prediction accuracy on the training set of 57 proteins.
[b]The rank of the observed interface against the generated surface patches, which have the same size as the interface, is divided into 10 equally sized categories for each protein and the ranks of the 57 proteins are averaged.
[c]The difference between the interface and the rest of the surface residues as calculated by individual or combined scores.
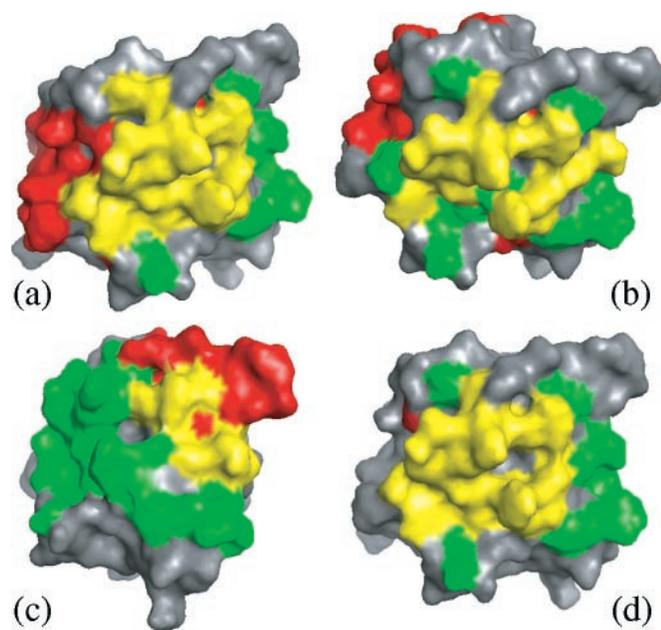

Figure 1. Comparison of protein interfaces predicted by single and combined scoring terms Red, predicted interface; green, observed interface; yellow, overlapped regions between predicted and observed interfaces. The interfaces were predicted by residue interface propensity (**a**), conservation score (**b**), energy score (**c**) and combination of them (**d**), respectively. There are actually two separate interfaces predicted by the conservation score. The small one is overlapped with the observed interface.

Dissection of interface properties

To understand the relative contribution of each scoring term in binding site prediction, the prediction accuracies of individual scoring terms and their pairwise combinations are shown in Table 3. Residue interface propensity is the most effective factor in predicting protein–protein interfaces with 37% coverage and 35% accuracy, respectively. Accuracy increases by 4% when interface propensity is combined with either energy score or conservation score and by another 6% when combined with both scores. All values are much higher than the average expected accuracy (15%). Figure 1 illustrates the interfaces predicted by the single and combined scoring terms for profilin (1pne). Different single terms give different levels of false positives (in red), true positives (in yellow) and false nagatives (in green) in interface prediction. The combined score gives the best prediction.

Table 4. The average side chain energies of interface, surrounding and other surface residues[a]

	Interface	Surrounding	Other surface
Mean energies(kcal·mol^{-1})	−0.79	−1.29	−1.18
SD	1.80	1.80	1.91
No. of residues	1271	1065	7579

[a]The surrounding residues have a lower energy than interface residues and other surface residues. The *P*-values are 5.9×10^{-12} and 0.038, respectively, as calculated by student's *t*-test. The surrounding residues are defined as those surface residues with side chain atoms within 1 Å plus sum of van der Waals radii of the two interacting atoms from any side chain in the interface.

One can also compare the average score of interfaces with that of surfaces. We find that only 31 out of the 57 proteins (54%) have more conserved interfaces. This weak-conservation result is consistent with findings of other groups (21,22). By comparison, the interface residues possess higher energy scores in 48 out of the 57 proteins. We also find that interface residues of 41 proteins (out of the 57 proteins) are more solvent accessible than other surface residues. We do not use surface accessibility as a term for binding prediction because a highly solvent-accessible residue usually has a higher energy score (38).

Another way to analyze the importance of individual terms is to test their abilities to rank observed interface in generated interfaces of the same size. As shown in Table 3, the energy score is the best in three terms for ranking observed interfaces. This happens despite that it has the lowest accuracy in interface prediction. The poorer performance of the energy score in prediction than in ranking is because high-energy interface residues are often surrounded by low-energy surface residues. As Table 4 shows, while the scores of interface residues are distinctively high, the scores of their surrounding residues are lower than those regular surface residues. Inclusion of non-interface surrounding residues leads to the low average rank (4.1) of the generated 20-residue initial patches that have maximum overlap (13 residues in average) with the observed binding interface, compared to 2.7 of the native interface.

Dependence on empirical parameters

There are many empirical parameters in the PINUP algorithm. For example, surface residues are defined as those

residues with >6% relative accessibility. We chose this small cut-off value for surface residues so that it is relative easy for the formation of a continuous path from the selected interface resides. If this cut-off value is set to 25%, we have to relax the definition of two connected residues in order to make possible the prediction of a continuous binding patch. Thus, we further define that two residues are connected when the distance between any two respective side chain atoms is <2 Å (rather than 1 Å) plus the sum of the van der Waals radii of the two atoms. The PINUP algorithm with the new cut-off values for surface residues and connected residues and new optimized weight factors yields an average coverage of 45%, accuracy of 44% and expected accuracy of 16%, respectively for the 57 proteins. Thus, the change of prediction accuracy (from 43, 45 and 15% for coverage, accuracy and expected accuracy, respectively) is small, compared to the change of cut-off values. Another important parameter is the size of the candidate surface patches. We set it to 20, a number close to the average size of the observed interfaces of the 57 proteins (22). If the patch size is set to 15 (25), the average prediction coverage and accuracy for the same 57 proteins are 40% (43%) and 40% (42%), respectively. Changes of other parameters are also tested. For example, selecting 10% of top ranked patches (rather than 5%) leads to 41% for coverage and 46% for accuracy. Selecting top 10 (or 20, default 15) residues as potential interface residues for small proteins leads to 38% (48%) for coverage and 46% (41%) for accuracy. All these calculations indicate that the overall accuracy of interface prediction has a weak dependence on empirical parameters. In general, if the coverage is higher, the accuracy is lower and vice versa. We have chosen the parameters so that coverage and accuracy are similar in magnitude.

Independent testing

PINUP is tested on the protein–protein docking bechchmark 2.0. The results for the whole set and a subset of proteins non-homologous (independent) to the training set are shown (Table 5). The average coverage, accuracy for the independent 68-protein set are 30.5 and 29.4%, respectively. They are 12 or 15% lower than the corresponding cross-validation value of the 57 training proteins.

The test set is further classified into rigid-body (easy), medium difficult and difficult sets according to the magnitude of conformational change after binding (45). Generally, the coverage and accuracy are reduced as the magnitude of binding-induced conformational changes increases (Table 5). Each subset is divided into enzymes, inhibitors and others. It is clear that most accurate predictions are made for inhibitors, as in the Jacknife cross validation of the 57-protein set. The average coverage and accuracy for 12 non-homologous inhibitors are 52.4 and 51.5%, respectively. The corresponding numbers are 26.5 and 26.6% for 7 enzymes, respectively and 25.7 and 24.4% for 49 other proteins, respectively. The excess accuracy over the expected accuracy is also the highest for inhibitors (27.2% for inhibitors versus 13.4% for enzymes and 12.5% for others).

The significantly reduced accuracy for the test set is not caused by over-training. The average coverage and accuracy (31.8 and 32.3%, respectively) is improved only a little even when the 68 proteins are used to train the two free parameters. The optimized weights for energy score, conservation score and propensity score are 1, 1.4 and 5, respectively. When the 57 proteins of Neuvirth *et al.* (10) are predicted with the new parameters, the coverage and accuracy are 43 and 44%, respectively.

The significantly reduced accuracy for the test set is also not caused by the significant homologous relation between the 75 proteins used for deriving interface propensity and the 57 proteins used for cross validation. We recalculate interface propensity by using the 68 protein complexes of the independent benchmark. The resulting propensity is used to predict the interfaces of the same 68 protein complexes. The prediction coverage and accuracy of the new method after re-optimizing the weight factors (31 and 32% for coverage and accuracy, respectively) are essentially unchanged from the results from independent testing (31 and 29%, respectively). This illustrates that interface propensity derived from the testing protein complexes did not help to improve the accuracy of interface prediction for those protein complexes.

We also recalculate interface propensity by using the 57 non-homologous protein complexes of the training and validation benchmark. For a given protein complex, it is excluded from the calculation of interface propensity and training of weight parameters. This jackknife cross validation yielded 40 and 41% for the average coverage and accuracy, respectively. (The average coverage and accuracy are 40 and 42%, respectively, when both interface propensity and

Table 5. Testing PINUP with the whole (in parentheses) and a subset of non-homologous proteins in the protein–protein docking benchmark 2.0[a]

Subset[b]	Category	No. of non-homologous proteins (all)[c]	Average (all) Coverage (%)	Accuracy (%)	Expected accuracy (%)
Rigid body	Enzyme	6 (21)	30.9 (42.1)	31.0 (50.5)	14.6 (14.3)
	Inhibitor	11 (21)	49.5 (60.4)	50.6 (57.4)	24.7 (25.7)
	Others	25 (44)	29.0 (31.7)	23.5 (29.5)	10.8 (13.1)
Medium difficult	Enzyme	1 (2)	0.0 (32.5)	0.0 (26)	5.1 (8.1)
	Inhibitor	1 (2)	84.6 (70.1)	61.1 (61.8)	20 (23.4)
	Others	16 (20)	25.0 (24.7)	28.1 (30.3)	12.1 (13.1)
Difficult	Others	8 (14)	16.9 (18.4)	20.0 (21.4)	14.7 (13.0)
All		68 (124)	30.5 (36.3)	29.4 (37.5)	14.2 (15.5)

[a]Benchmark 2.0 of Chen *et al.* (45) except antibody–antigen complexes.
[b]Subset is based on the magnitude of conformational change after binding (45).
[c]The number of proteins that share sequence identity <35% with any protein in the 57-protein dataset compiled by Neuvirth *et al.* (10). The number in parentheses is averaged for all proteins in the category.

S_r^{ave} are derived from the training and cross validation set). This is only 3% (or 4%) lower than the average coverage (or accuracy) predicted from the original interface propensity. The result further confirms that the homologous proteins used for calculation of interface propensity is not responsible for the higher accuracy for the training and cross-validation benchmark.

Lower prediction accuracy for the test benchmark means that the binding sites of proteins in the training benchmark are easier to predict than that in the testing benchmark. This is because the binding interfaces for the proteins in the testing benchmark are less distinguishable from the rest of protein surfaces than in the training benchmark. Indeed, the difference of the interface propensity between 20 amino acid residues derived from the independent benchmark of 68 proteins is significantly smaller than that of the original interface propensity. Moreover, interfaces regions are more conserved than surface regions only in 26 of the 68 proteins (38%), compared to 54% for the 57-protein training set. The difficulty to predict interfaces of some proteins are also found in other studies (36,37). The varying performance in different benchmarks indicates the limitation of existing benchmarks. This is the direct consequence of a limited number of NR transient complex structures available in the protein data-bank. Thus, it would be more representative to combine two benchmarks into one. Because 56 proteins (or their homologs) in the 57 protein training set are in the protein–protein docking benchmark 2.0, the results for all 124 proteins in this benchmark (36.3 coverage and 37.5% for accuracy, see Table 5) provide a more reasonable estimate of the overall accuracy for PINUP.

Comparison with other methods

It is difficult to compare performance made by different methods for binding site prediction. Some (10,35) optimize their methods to make a few accurate predictions (high accuracy and low coverage) while others (36) emphasize on a large coverage of actual binding site (high coverage). More importantly, different definitions of surface/interface residues also produce different expected values from random prediction. For example, a relaxed definition of interface residues will lead to an increase in expected accuracy. Thus, the comparison between different methods has to focus on the excess from expected values. Here, we make an attempt to have a comparison between PINUP and two recent studies.

Neuvirth *et al.* (10) developed a method called ProMate to distinguish interface regions from the rest of protein surfaces based on 13 properties. For the same 57 proteins for training and cross validation, they reported a success rate of 36 proteins with accuracy of 50% or higher. This rate is significantly higher than a random success rate of 13 proteins. By comparison, the corresponding success rate and expected success rate for PINUP are 25 and 3 proteins, respectively. Thus, the excess success rate from the expected value given by PINUP (22) is similar to that given by ProMate (23). However, PINUP gives a significantly higher number of proteins with coverage of 50% or higher than ProMate does. It is 20 for PINUP but 0 for ProMate. Here, we used the same random model proposed by Neuvirth *et al.* to generate expected values. In this model, the predicted scores are reshuffled over

the amino acid residues before extracting the patches. The low rate of correct prediction in application of this random model to the PINUP algorithm is mostly due to the relatively large number of predicted interface residues in this study. The larger the number of predicted residues, the smaller the random probability that half of the predicted residues are true interface residues. By comparison, ProMate focuses on making a small number of predictions and thus intrinsically has a high expected success rate. Thus, PINUP is as accurate as ProMate in terms of the location of interface but has a larger overlap with the actual interface.

Chen and Zhou (37) developed a method called PPISP, which trained a neural network with sequence profiles and solvent accessibility of spatially neighboring surface residues to predict interface residues. In their paper, they defined coverage and accuracy as the fraction of correctly predicted interface residues for all proteins in total observed interface residues for all proteins and in total predicted interface residues for all proteins, respectively. Similarly, the expected accuracy was defined as the fraction of total interface residues in total surface residues for all proteins. To distinguish from the coverage and accuracy defined in this work, we called coverage and accuracy defined by Chen and Zhou as the overall coverage and overall accuracy, respectively. Their method was tested on protein–protein docking benchmark 1.0 of Chen *et al.* (45) The overall coverage and overall accuracy is 50% and 50% for the enzyme-inhibitor category but only 28 and 31%, respectively, for other proteins in the rigid body group. PINUP is tested on the more extensive benchmark 2.0. Results are shown in Table 5. PINUP yields 49% of overall coverage and 52% of overall accuracy for the enzyme-inhibitor category. The corresponding values are 29 and 27%, respectively, for others. Thus, the overall accuracy is similar between PINUP and PPISP. However, the overall expected accuracy (13% for enzymes, 20% for inhibitors and 10% for others) of PINUP for this benchmark may be lower than that of PPISP (26% for all heterodimers in the training set).

DISCUSSION

In this paper, we present the method PINUP to locate binding sites of unbound structures of proteins. This method is based on a scoring function that is a linear combination of a side-chain energy score, interface propensity and residue conservation score. We find that the linear combination provides a 10% improvement in accuracy and 6% improvement in coverage over the residue interface propensity—the best single-term score function. Thus, combining independent parameters is one reason for achieving high coverage and high accuracy at the same time.

The combination of residue-energy and conservation scores also serves the purpose of separation of functional constraints and structural constraints for binding site predictions. A residue with a low energy is likely conserved for structural constraints. Subtraction of the negative energy score (expected conservation score for structural reasons) from the observed conservation score likely produces a score to better identify functionally important residues. In the work of Chelliah *et al.* structural constraints are separated from functional

constraints by using surface accessibility (30). Here we use a residue energy score because the energy score is more effective than solvent accessibility in distinguishing interface regions from the rest of protein surface (38).

Another reason for the improved prediction of PINUP is to define interface as the largest continuous patches composed of residues having highest appearance in 5% top-scored initial patches, rather than as the top-scored initial patch. This consensus over top-scored initial patches permits the prediction of a patch with varied shape and size whereas the initial patch is round and fixed at 20 residues. More importantly, the interfaces in large unbound monomers are often made of relatively buried, low-energy residues surrounded by high-energy residues (38). As a result, a single top-scoring patch centered on a peripheral residue will contain only a small number of the interface residues. Indeed, if only one top-scored initial patch is selected, the coverage and accuracy are only 38 and 39%, respectively, for the set of 57 proteins.

The improvement in binding-site prediction is also in part due to the use of native coordinates. Side-chain energy score, originally developed for protein design is a weighted average energy score over all possible rotamers. Unlike the early work (38), the coordinates of the rotamer that is closest to the native side chain conformation are replaced by the coordinates of native side-chain to reduce the discrete error. Without this replacement, the combined score would produce 42% of accuracy rather than 45% (Table 3).

Recently, Cheng *et al.* found that functional sites are more likely (than non-functional sites) to have computed sequence profiles that differ significantly from the naturally occurring sequence profiles and to have residues with sub-optimal free energies (31). We have tested a new energy score $\ln(n + 1)$, where n is the number of other residue types having lower energies than the replaced native residue. This single residue-energy score yields a accuracy of 24% and its combination with conservation score yields a accuracy of 33%. This performance is much better than the original energy score. However, the combination of this new score with conservation and interface propensity leads to an accuracy of 43%. It is still slightly lower than 45% by PINUP. One possible reason is the stronger correlation between the new energy score and interface propensity. Residues with a high-energy score of the new energy term often are exposed hydrophobic residues, which also have high interface propensity score. In addition, the redefined energy score cannot be calculated with the native side-chain conformation. The new energy scores of all residue types were calculated with rotamers because there is no native conformation for mutants and this discretization error may also affect the prediction accuracy.

We also test a method that defines interface residues based on the scores of individual residues. The levels of accuracy for predicted interfaces are 22, 21 and 24% for the top 15% surface residues by using energy score, conservation score and interface propensity, respectively. Compared to the corresponding values in Table 3, the use of patch is useful for a more accurate prediction of binding sites for conservation score and interface propensity but not for the energy score. This further highlights the negative effect of high-energy interface residues surrounded by low-energy surface residues on the ability of the energy score to rank initial patches.

One perhaps notes that some false positives predicted by PINUP are actually the binding site of other molecules instead of the expected partner protein. This is because proteins often have multiple interacting partners. Some bind to ligands and others interact with other proteins or nucleic acids. The predicted protein binding site of aspartate transcarbamoylase (1ekx) is overlapped with the PAL ligand binding site, rather than the observed protein–protein interface. A domain–domain interaction site of Enteropathogenic *E.coli* intimin C-terminal domain (1f00), is predicted as the protein–protein binding interface. Thus, in future, we plan to develop a significance score for predicted interface regions. Those with a high significance score (rather than only one interface) will be treated as predicted interfaces. This will allow the possibility of predicting multiple interface regions for a given protein.

One interesting observation is that PINUP provides a much more accurate prediction for interfaces of inhibitors. This is true for both training and testing sets. Similar results were observed in the previous study, where interface residues were predicted by merely energy score (38). One possible reason is that inhibitors have a higher expected accuracy. That is, it has a larger interface in a relatively small surface, compared to other proteins.

ACKNOWLEDGEMENTS

REFERENCES

1. Bork,P., Jensen,L.J., von Mering,C., Ramani,A.K., Lee,I. and Marcotte,E.M. (2004) Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.*, **14**, 292–299.
2. Aloy,P. and Russell,R.B. (2006) Structural systems biology: modelling protein interactions. *Nature Rev. Mol. Cell. Biol.*, **7**, 188–197.
3. Wodak,S.J. and Mendez,R. (2004) Prediction of protein–protein interactions: the CAPRI experiment, its evaluation and implications. *Curr. Opin. Struct. Biol.*, **14**, 242–249.
4. Mendez,R., Leplae,R., De Maria,L. and Wodak,S.J. (2003) Assessment of blind predictions of protein–protein interactions: current status of docking methods. *Proteins*, **52**, 51–67.
5. Liu,S., Zhang,C., Zhou,H. and Zhou,Y. (2004) A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins*, **56**, 93–101.
6. Aloy,P., Querol,E., Aviles,F.X. and Sternberg,M.J. (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.*, **311**, 395–408.
7. Russell,R.B., Alber,F., Aloy,P., Davis,F.P., Korkin,D., Pichaud,M., Topf,M. and Sali,A. (2004) A structural perspective on protein–protein interactions. *Curr. Opin. Struct. Biol.*, **14**, 313–324.
8. Zhang,C., Liu,S. and Zhou,Y. (2005) Docking prediction using biological information, ZDOCK sampling technique, and clustering guided by the DFIRE statistical energy function. *Proteins*, **60**, 314–318.
9. Jones,S. and Thornton,J.M. (1997) Prediction of protein–protein interaction sites using patch analysis. *J. Mol. Biol.*, **272**, 133–143.

10. Neuvirth,H., Raz,R. and Schreiber,G. (2004) ProMate: a structure based prediction program to identify the location of protein–protein binding sites. *J. Mol. Biol.*, **338**, 181–199.

11. Glaser,F., Steinberg,D.M., Vakser,I.A. and Ben-Tal,N. (2001) Residue frequencies and pairing preferences at protein–protein interfaces. *Proteins*, **43**, 89–102.

12. Young,L., Jernigan,R.L. and Covell,D.G. (1994) A role for surface hydrophobicity in protein–protein recognition. *Protein Sci.*, **3**, 717–729.

13. Lo Conte,L., Chothia,C. and Janin,J. (1999) The atomic structure of protein–protein recognition sites. *J. Mol. Biol.*, **285**, 2177–2198.

14. Jones,S. and Thornton,J.M. (1996) Principles of protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.

15. Zhou,H.X. and Shan,Y. (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins*, **44**, 336–343.

16. Fariselli,P., Pazos,F., Valencia,A. and Casadio,R. (2002) Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem.*, **269**, 1356–1361.

17. Panchenko,A.R., Kondrashov,F. and Bryant,S. (2004) Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci.*, **13**, 884–892.

18. Pupko,T., Bell,R.E., Mayrose,I., Glaser,F. and Ben-Tal,N. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18**, S71–S77.

19. Chung,J.L., Wang,W. and Bourne,P.E. (2006) Exploiting sequence and structure homologs to identify protein–protein binding sites. *Proteins*, **62**, 630–640.

20. Hu,Z., Ma,B., Wolfson,H. and Nussinov,R. (2000) Conservation of polar residues as hot spots at protein interfaces. *Proteins*, **39**, 331–342.

21. Bradford,J.R. and Westhead,D.R. (2003) Asymmetric mutation rates at enzyme-inhibitor interfaces: implications for the protein-protein docking problem. *Protein Sci.*, **12**, 2099–2103.

22. Caffrey,D.R., Somaroo,S., Hughes,J.D., Mintseris,J. and Huang,E.S. (2004) Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.*, **13**, 190–202.

23. Grishin,N.V. and Phillips,M.A. (1994) The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. *Protein Sci.*, **3**, 2455–2458.

24. Mintseris,J. and Weng,Z. (2005) Structure, function, and evolution of transient and obligate protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **102**, 10930–10935.

25. Lichtarge,O., Bourne,H.R. and Cohen,F.E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.

26. Madabushi,S., Yao,H., Marsh,M., Kristensen,D.M., Philippi,A., Sowa,M.E. and Lichtarge,O. (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.*, **316**, 139–154.

27. Herzberg,O. and Moult,J. (1991) Analysis of the steric strain in the polypeptide backbone of protein molecules. *Proteins*, **11**, 223–229.

28. Elcock,A.H. (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.*, **312**, 885–896.

29. Ota,M., Kinoshita,K. and Nishikawa,K. (2003) Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *J. Mol. Biol.*, **327**, 1053–1064.

30. Chelliah,V., Chen,L., Blundell,T.L. and Lovell,S.C. (2004) Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *J. Mol. Biol.*, **342**, 1487–1504.

31. Cheng,G., Qian,B., Samudrala,R. and Baker,D. (2005) Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucleic Acids Res.*, **33**, 5861–5867.

32. Jones,S. and Thornton,J.M. (1997) Analysis of protein–protein interaction sites using surface patches. *J. Mol. Biol.*, **272**, 121–132.

33. Cole,C. and Warwicker,J. (2002) Side-chain conformational entropy at protein–protein interfaces. *Protein Sci.*, **11**, 2860–2870.

34. Fernandez,A. and Scheraga,H.A. (2003) Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proc. Natl Acad. Sci. USA*, **100**, 113–118.

35. Bradford,J.R. and Westhead,D.R. (2005) Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics*, **21**, 1487–1494.

36. Bordner,A.J. and Abagyan,R. (2005) Statistical analysis and prediction of protein–protein interfaces. *Proteins*, **60**, 353–366.

37. Chen,H. and Zhou,H.X. (2005) Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data. *Proteins*, **61**, 21–35.

38. Liang,S., Zhang,J., Zhang,S. and Guo,H. (2004) Prediction of the interaction site on the surface of an isolated protein structure by analysis of side chain energy scores. *Proteins*, **57**, 548–557.

39. Liang,S. and Grishin,N.V. (2004) Effective scoring function for protein sequence design. *Proteins*, **54**, 271–281.

40. Dunbrack,R.L.,Jr and Cohen,F.E. (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.*, **6**, 1661–1681.

41. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

42. Brooks,B.R., Bruccoleri,R.E., Olafson,B.D., States,D.J., Swaminathan,S. and Karplus,M. (1983) CHARMM: a program for macromolecular energy, minimization and dynamics calculation. *J. Comput. Chem.*, **4**, 187–217.

43. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

44. Word,J.M., Lovell,S.C., Richardson,J.S. and Richardson,D.C. (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, **285**, 1735–1747.

45. Chen,R., Mintseris,J., Janin,J. and Weng,Z. (2003) A protein–protein docking benchmark. *Proteins*, **52**, 88–91.