Prediction of the interaction site on the surface of an isolated protein structure by analysis of side chain energy scores

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · NOVEMBER 2004

Impact Factor: 2.63 · DOI: 10.1002/prot.20238 · Source: PubMed

4 AUTHORS, INCLUDING:

Shicui Zhang
Ocean University of China
217 PUBLICATIONS **2,461** CITATIONS
SEE PROFILE

Huarong Guo
Ocean University of China
21 PUBLICATIONS **154** CITATIONS
SEE PROFILE

PROTEINS: Structure, Function, and Bioinformatics 57:548–557 (2004)

Prediction of the Interaction Site on the Surface of an Isolated Protein Structure by Analysis of Side Chain Energy Scores

Shide Liang,* Jian Zhang, Shicui Zhang, and Huarong Guo
Department of Marine Biology, Ocean University of China, Qingdao, People's Republic of China

ABSTRACT **We show that residues at the interfaces of protein–protein complexes have higher side-chain energy than other surface residues. Eight different sets of protein complexes were analyzed. For each protein pair, the complex structure was used to identify the interface residues in the unbound monomer structures. Side-chain energy was calculated for each surface residue in the unbound monomer using our previously developed scoring function.[1] The mean energy was calculated for the interface residues and the other surface residues. In 15 of the 16 monomers, the mean energy of the interface residues was higher than that of other surface residues. By decomposing the scoring function, we found that the energy term of the buried surface area of non-hydrogen-bonded hydrophilic atoms is the most important factor contributing to the high energy of the interface regions. In spite of lacking hydrophilic residues, the interface regions were found to be rich in buried non-hydrogen-bonded hydrophilic atoms. Although the calculation results could be affected by the inaccuracy of the scoring function, patch analysis of side-chain energy on the surface of an isolated protein may be helpful in identifying the possible protein–protein interface. A patch was defined as 20 residues surrounding the central residue on the protein surface, and patch energy was calculated as the mean value of the side-chain energy of all residues in the patch. In 12 of the studied monomers, the patch with the highest energy overlaps with the observed interface. The results are more remarkable when only three residues with the highest energy in a patch are averaged to derive the patch energy. All three highest-energy residues of the top energy patch belong to interfacial residues in four of the eight small protomers. We also found that the residue with the highest energy score on the surface of a small protomer is very possibly the key interaction residue. Proteins 2004;57:548–557.**

Key words: protein–protein interface; side-chain energy; binding free energy; patch analysis; molecular recognition

INTRODUCTION

Protein–protein interactions play a key role in many biological processes. The structure of a protein complex is generally more difficult to obtain than the monomer structures. The goal of protein docking is to obtain a model for the bound complex from the atomic coordinates, determined by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy of the unbound component molecules.[2,3] Current docking methods usually search the six-dimensional space of relative orientations of the two molecules and identify a set of candidate complex structures by evaluation of shape complementarity.[4] These structures are then rescored according to some force field or statistical residue–residue contact potential.[5–7] After two steps of scoring, many false positives still exist in addition to near-native states, mostly due to shape variability of the interacted proteins during molecular association.[7,8] One approach to eliminating the false positives involves examining the chemical characteristics and residue propensities of the interfacial region, which are not sensitive to small conformational changes, and re-ranking the complex structures. In more difficult cases, the experimental structure is available for only one of the two interacted proteins. The putative binding site has to be identified by selecting the patch with specific features on the surface of an isolated protein structure.[9,10]

Native protein binding sites have exceptional attributes, which are different from those of the rest of the protein surface. The interface sites are usually poorer in polar/charged residues and richer in hydrophobic residues than the average protein surface, which indicates that the hydrophobic effect plays a key role in protein–protein association.[11–13] A similar observation has been made that interfaces are much richer in the aromatic residues His, Tyr, Phe and Trp than surfaces (21% vs. 8 %) and somewhat richer in the aliphatic residues Leu, Ile, Val and Met (17% vs. 11%).[14] They are depleted in the charged residues Asp, Glu and Lys but not in Arg, which is the residue type that makes the largest overall contribution (10%) to interfaces. Cole and Warwicker have estimated side-chain conformational entropy (per unit solvent acces-

Abbreviations: PDB, protein data bank; OMTKY3, turkey ovomucoid third domain; BPTI: bovine pancreatic trypsin inhibitor

*Correspondence to: S. Liang, Department of Marine Biology, Ocean University of China, Qingdao, 266003, People's Republic of China. E-mail: shideliang@hotmail.com.

Received 11 January 2004; Revised 5 March 2004, 4 May 2004; Accepted 6 May 2004

Published online 29 July 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20238

sible area) for various protein surface regions, using a self-consistent mean field calculation of rotamer probabilities.[15] Interfacial surface regions were found to be less flexible than the rest of the protein surface in 21 of 25 cases for the monomer extracted from homodimer datasets and in eight of nine cases for the large protomer extracted from heterodimer datasets. More favorable binding surfaces are expected to correlate with a smaller loss of side-chain conformational entropy and a larger burial of nonpolar surface area. In another study, most backbone hydrogen bonds in the majority of soluble proteins were found to be thoroughly wrapped intramolecularly by nonpolar groups except for a few likely to surround the binding site.[16] The underwrapped hydrogen bonds may be dramatically stabilized by the removal of water during molecular association.

Jones and Thornton have calculated six parameters–solvation potential, residue interface propensity, hydrophobicity, planarity, protrusion and residue accessible surface area–for the observed interface patch and all other surface patches defined on each sampled protein.[9,10] Although the interfaces are among the most planar and accessible patches, other properties differ according to the nature of the protein complex. Interfaces of homodimers are more hydrophobic than those of hetero-complexes, which can be among the most polar patches on the surface. Since hydrophilic atoms at the interfaces are already partially desolvated before binding,[17] hydrogen bonds and electrostatic interactions, in addition to hydrophobic interactions, may contribute favorably to the binding free energy. The favorable interactions between a hydrophilic pair could be larger than the cost of further desolvation of the two hydrophilic atoms. In other words, the buried non-hydrogen-bonded hydrophilic atoms or the exposed hydrophobic residues are in high-energy states. Interfacial residues of any type could have a higher energy score than other surface residues in the unbound structure, and the 'sticky' interfacial regions of native proteins are designed to interact with other proteins.[18] However, this idea has not been confirmed by computational analysis of the energy of residues on the protein surface.

We have developed a scoring function for protein design.[1] The scoring function is a linear combination of the following terms: (1) atom contact surface and overlap volume, (2) hydrogen bond energy, (3) electrostatic interactions, (4) desolvation energy, (5) rotamer intrinsic energy, (6) disulfide bond energy, (7) reference value of each amino acid. The weights of these energy items, along with the reference values, are optimized so that the native residue is predicted to be energetically favorable at each position of the training proteins. The success of our scoring function has been demonstrated by predicting mutant changes in stability for the tested proteins. In this study, we used the scoring function to evaluate side-chain energy scores of surface residues for the unbound monomers of protein complexes. We found that interfacial residues have higher energy than other surface residues. Patch analysis of side-chain energy on the surface of an isolated protein structure could be useful in identifying the putative protein-binding site.

MATERIALS AND METHODS

Protein Structures

Eight protein–protein complexes representing four major classes were analyzed: the barnase–barstar (PDB code 1brs) system from the RNAase-inhibitor family, hen egg white lysozyme bound to antibody Fab D44.1 (1mlc) from the antigen-antibody family, acetylcholinesterase–fasciculin II (1fss) from the serine esterase-toxin family and five systems from the proteinase-inhibitor family, including human leukocyte elastase–OMTKY3 (1ppf), chymotrypsin–OMTKY3 (1cho), trypsin–BPTI (2ptc), subtilisin BPN′–streptomyces subtilisin inhibitor (2sic) and subtilisin novo-chymotrypsin inhibitor 2 (2sni). The same protein sets have been studied by Camacho and Vajda.[2] Both the complex and the unbound monomer structures of each protein set are available in the PDB.[19] The program REDUCE[20] was used to add hydrogen atoms to all proteins. Non-polar hydrogen atoms and all water molecules were deleted.

Scoring Function Used to Calculate Side-Chain Energy

The scoring function used to calculate the energy of a rotamer,[21] the representative conformation of the amino acid placed on the modeled position, was obtained in a previous study:[1]

$$E = -0.143 \times S_{\text{contacr}} + 0.724 \times V_{\text{overlap}} + 1.72 \times E_{\text{hbond}} + 28.6 \times E_{\text{elec}} - 0.0467 \times \Delta S_{\text{pho}} + 0.0042 \times \Delta S_{\text{phi}} + 1.14 \times \Delta(F_{\text{phi}})^{30} + 7.95 \times V_{\text{exclusion}} - 0.919 \times \ln(f_1 \times f_2) - 4.3 \times N_{\text{ssbond}} - \Delta G_{\text{ref}} \quad (1)$$

where S_{contact}, V_{overlap}, E_{hbond}, E_{elec}, ΔS_{pho} and ΔS_{phi} are atom contact surface, overlap volume, hydrogen bonding energy, electrostatic interaction energy, buried hydrophobic solvent accessible surface and buried hydrophilic solvent accessible surface between the rotamer and other parts of the protein, respectively; F_{phi} is the fraction of 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 given a backbone conformation; N_{ssbond} is the flag of disulfide bridge (1 or 0); ΔG_{ref} is assumed to be the difference between the free energy of the rotamer in solvent and denatured protein.

The weights of the energy terms and the reference value of each amino acid were determined by minimizing the sum of the following equation over 5792 positions of the training set of 28 proteins:

$$-\ln\frac{\sum\exp(-E(r_{\text{native}}))}{\sum\exp(-E(r_{\text{all}}))} \quad (2)$$

TABLE I. Comparison of Side-Chain Energy Between Interfacial Residues and Other Surface Residues

Unbound Monomers	PDB Code	Complex Code	Mean Energy Score of Residues (kcal/mol)			No. of Patches	Size of Interface	Rank of Interface
			Interface	Non-Interface	Difference			
Elastase	1ppg	1ppf	−0.67	−0.98	0.31	139	21	13
OMTKY3[a]	2ovo		−0.29	−1.28	0.99	52	14	1
α-Chymotrypsin	5cha	1cho	−0.78	−1.02	0.24	180	22	50
OMTKY3	2ovo		0.04	−1.37	1.41	52	13	1
Acetylcholinesterase	2ace	1fss	−0.79	−1.15	0.36	343	25	62
Fasciculin-II	1fsc		−0.78	−1.32	0.54	56	20	1
Barnase	1a2p	1brs	−1.21	−1.29	0.08	90	18	34
Barstar	1a19		−0.42	−0.95	0.53	70	17	7
Subtilisin BPN′	2st1	2sic	−0.90	−1.30	0.40	165	25	35
Streptomyces inh.	3ssi		−0.03	−0.92	0.89	92	13	1
Trypsin	2ptn	2ptc	−1.05	−1.30	0.24	155	23	26
BPTI[b]	6pti		−0.08	−1.16	1.08	44	11	1
Subtilisin novo	2sbt	2sni	0.29	0.30	−0.02	202	28	105
Chymotrypsin inh.2	2ci2		−0.29	−0.93	0.65	56	15	4
Fab D44.1	1mlb	1mlc	−0.51	−0.78	0.27	331	19	61
Hen egg lysozyme	1lza		−0.98	−1.50	0.52	97	17	10
Mean			−0.53	−1.06	0.53			

where r_{native} is a rotamer of the native residue type at the modeled position and $E(r_{\text{native}})$ is energy of the rotamer; the summation space in the numerator is over rotamers of the native residue type, while the partition function in the denominator is over all rotamers of 20 amino acids. The free energy of a particular amino acid on the modeled position was calculated as

$$-\ln(\sum\exp(-E(r_i))) \quad (3)$$

where r_i is a rotamer of the amino acid and i is the index of the rotamers. The value calculated using eq. (3) was then divided by 2.41, the slope of the regression line between the calculated and experimental unfolding $\Delta\Delta G$ of a set of point mutation data.[1] Thus the energy unit can be set to the same value as the experimental data (kcal/mol).

Definition of Observed Interface and Surface Patches

Surface residues were defined as those residues with a side-chain that has a relative accessibility >6%. The C_α atom of glycine was considered to be a side-chain atom for convenience of calculation. Atomic accessible surface was calculated in the same way used in the previous study.[1] The solvent probe was set to 1.2 Å. Since we did not calculate interactions between hetero-atoms and amino acid atoms, residues with any side-chain atom neighboring hetero-atoms at a distance of less than 6 Å were excluded in the analysis. The observed interface was defined as the set of qualified surface residues with side-chains possessing accessible surface areas that decrease by >1 Å upon complexation. A patch was defined as a central surface accessible residue and n nearest surface accessible neighbors (n + 1 is the size of the patch in term of the number of the residues), as defined by C_α positions. Solvent vector[9] constraints were applied in order to avoid patches forming rings around the surface, or patches sampling through the center of a protein. The solvent vector of a surface residue was defined as the vector from the barycenter of the nearest ten C_α atoms of this residue to the C_α atom of itself. The angles between the solvent vectors of the central residue and each of the surrounding residues in the patch were <110°.

RESULTS AND DISCUSSION

Analysis of Side-Chain Energy at Protein–Protein Interfaces

Eight different sets of protein complexes were analyzed. For each protein pair, the complex structure was used to identify the interfacial residues in the unbound monomer structures. Side-chain energy was calculated for each surface residue in the unbound monomer, and the mean value was calculated for the interfacial residues and the rest of the surface residues. In 15 of the 16 monomers, the mean value of side-chain energy of interface residues was higher than that of other surface residues. The mean energy of the interface residues of subtilisin novo was only slightly lower than that of other surface residues. We then ranked the energy of the observed interface relative to the generated surface patches for each monomer. The patches were defined as clusters of residues surrounding the central residue on the protein surface. They were overlapped and approximately circular in shape. Each surface residue was qualified as the central residue, and the number of the generated patches was equal to the number of surface residues. For an individual protein, the size of the surface patch was equal to the observed interface in terms of residue number. Table I lists the ranks of the observed interface patches of the small and large protomers for each protein pair. The ranks of the small protomers are much higher than those of the large protomers. In fact, the interface patches of five small promoters rank highest among the surface patches.

The interface patches of the large protomers rank lower than those of small protomers because interface residues

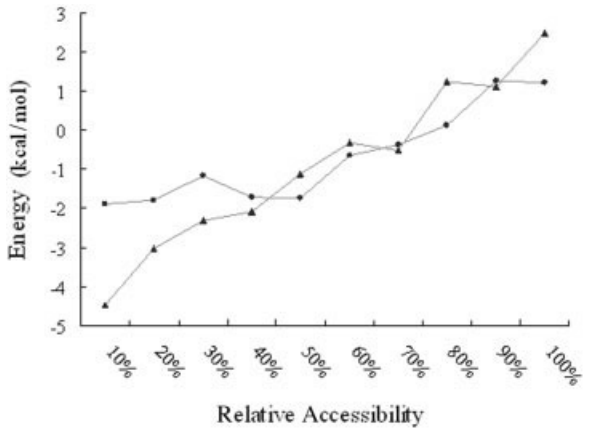

Fig. 1. Distributions of the side-chain energy of residues with variable accessibilities. ▲, Leu; ●, Asp. The unbound monomers of the protein complexes were used for the statistical analysis.

of the large protomers are less solvent accessible. The interfaces of large protomers are usually concave, while those of small protomers can be quite convex. Residues at the concave surface are usually less solvent accessible than those at the convex surface. When residue relative accessibility was used as the evaluation method, the mean value of the decile ranks was 1.5 for the eight small protomers and 7 for the eight large protomers. Here, a rank of 1 would indicate that the relative accessibility of the interface patch falls in the highest 10% range of the distribution of all surface patches. Our scoring function predicts the buried residues, which form more interactions with other residues than the surface residues, to be low energy whether they are hydrophobic or hydrophilic (Fig. 1). The buried hydrophilic residues usually form hydrogen bonds or salt bridges in native proteins. The desolvation cost of the hydrophilic atoms is compensated by hydrogen bonding and electrostatic interaction energy. The hydrophobic residues, which have low reference values,[1] are predicted to receive low energy scores only if they are deeply buried and form favorable hydrophobic interactions. In fact, the low energy of the buried residues is mostly due to their large contact surfaces with other residues. It is worth noting that the energy values are not simply determined by the solvent accessibility. The interface residues are less solvent accessible than other surface residues in six of the eight large protomers. However, all six large protomers have higher side-chain energy at the interfaces than in other surface regions. Obviously, other factors beyond the relative accessibility contribute to the high energy of the interface patches.

Prediction of Protein–Protein Interfaces by Patch Energy Analysis

Identification of the putative interaction site on the surface of an isolated protein is a challenging task in the field of molecular recognition. To avoid combinatorial searches, only simply defined surface patches (see Materials and Methods) were sampled in this study. The patch size was set to 20 for all kinds of monomers. Percentage

TABLE II. Prediction of Protein-Protein Interfaces by Patch Energy Analysis

Proteins	Overlap of Max Patch[a]	Rank of Max Patch[b]	Overlap of Top Patch[c]
Elastase	70.0	21	35.0
OMTKY3	92.9	6	71.4
α-Chymotrypsin	70.0	39	0.0
OMTKY3	100.0	6	76.9
Acetylcholinesterase	80.0	65	40.0
Fasciculin-II	75.0	30	55.0
Barnase	72.2	54	33.3
Barstar	94.1	18	82.4
Subtilisin BPN′	90.0	10	30.0
Streptomyces inh.	100.0	1	100.0
Trypsin	75.0	43	30.0
BPTI	100.0	5	0.0
Subtilisin novo	75.0	107	0.0
Chymotrypsin inh.2	93.3	27	20.0
Fab D44.1	63.2	18	0.0
Hen egg lysozyme	94.1	6	70.6

[a]Overlap value of the patch which had the maximum overlap value with the observed interface.
[b]Rank of the patch with the maximum overlap value.
[c]Overlap value of the patch with top energy.

overlap of the predicted patch with the observed interface was calculated as:

$$\frac{N_{\text{overlap}}}{\text{Min}(N_{\text{interface}}, N_{\text{patch}})} \times 100 \quad (4)$$

where N_{overlap} is the number of residues present both at the observed interface and in the predicted patch, $N_{\text{interface}}$ is the number of residues at the observed interface and N_{patch} is the patch size. Table II lists the results. Six of the 16 predicted interfaces have an overlap of >50% with the true interface, and 11 of the 16 have an overlap of ≥30%. The interface patch of streptomyces subtilisin inhibitor was correctly predicted (i.e. the patch, which has the maximum overlap with the observed interface, was calculated as the highest energy score among all sampled surface patches). The prediction results were affected by the use of a circular surface patch and the artificially set patch size. It is unlikely that one of the sampled patches on the protein surface would exactly match the observed interface.

None of residues of the top energy patch overlaps with the observed interface in four monomers (chymotrypsin, BPTI, subtilisin novo, and antibody Fab D44.1). BPTI has the smallest interface size of 11 out of the 16 monomers, while subtilisin novo has the largest interface size of 28. If the patch size were reset to the same values as the respective interfaces, the top energy patch would be 64% overlapped with the observed interface for BPTI and 32% overlapped for subtilisin novo. The interface sizes of chymotrypsin and antibody Fab D44.1 are 22 and 19 respectively, very similar to the artificially set patch size of 20. Resetting the patch size has little effect on the predicted results. According to visual analysis, the interface of chymotrypsin is concave and that of Fab D44.1 is quite

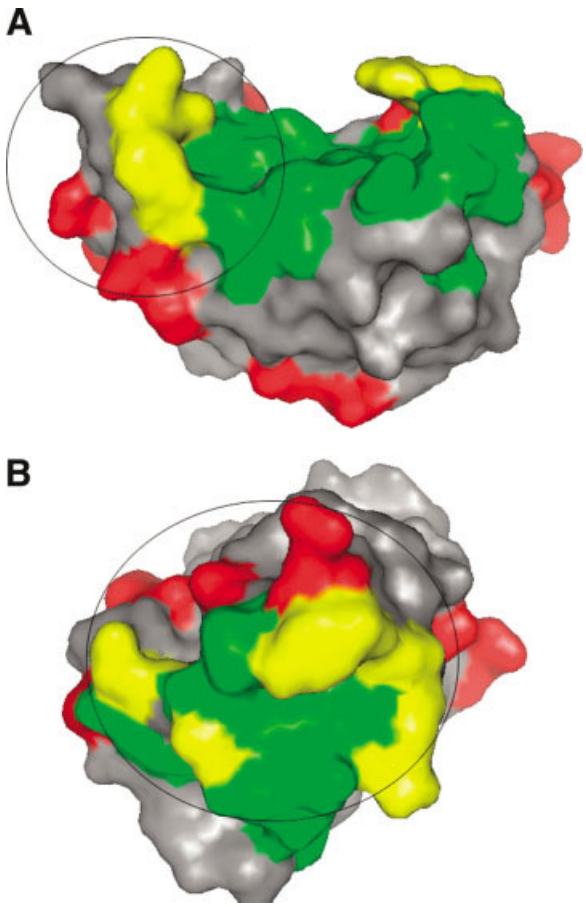

Fig. 2. The side-chain energy scores of residues at the barnase–barstar interface. Red, residues with the side chain energy in the range of top 20%; green, interface residues with the side chain energy in the top 20%; circle, the surface patch with top energy. (**A**) Barnase; (**B**) Barstar. Structures were rendered in Pymol (DeLano Scientific, San Carlos, CA).

flat. The decile ranks of the interface patches are eighth and ninth, respectively in terms of residue accessibility. The failure of predicted results is partly due to the relatively buried interface residues. In addition, the 13 highest-energy patches in antibody Fab D44.1 are located on the tip of the C_{H1} domain that would contact the C_{H2} domain in an intact antibody molecule. These patches overlap. As indicated by other studies,[10,15] additional protein–protein or domain–domain interaction sites could result in failed predictions. Some residues in this region are poorly resolved in X-ray diffraction, and the atomic positions should be considered arbitrary,[22] which also contributes to the incorrect predictions. In fact, the 14^{th} highest-energy patch in antibody Fab D44.1 was found to overlap with the observed interface.

The prediction accuracy was found to be dependent on the shape of the interface. The interfaces of the large protomers are concave and are usually composed of protruding peripheral residues and sunken central residues. Our scoring function calculated the peripheral residues as having high energy and the relatively buried central residues as having low energy. The patch with the top energy could cover part of the peripheral residues and avoid the central residues [Fig. 2(a)]. Therefore the overlap value between the top energy patch and the observed interface is low for the large protomers (Table II). The situation is different for the small protomers. The interface residues of the small protomers are located on the convex surfaces. There is no longer a low energy region with large size at the center of the interfaces. The top energy patch can cover most of the interfacial residues [Fig. 2(b)].

Residues with Highest Energy in Surface Patches

In the six small protomers (OMTKY3/elastase, OMTKY3/chymotrypsin, fasciculin-II, barstar, streptomyces inh. and BPTI), the residue with the highest energy score on the protein surface is located at the interface. The residue with the second highest energy is located at the interface in the other two small protomers (chymotrypsin inh.2 and hen egg lysozyme). Searching for the residue with the highest energy therefore provided an easy way to predict the location of the protein–protein interface. However, a single residue is not enough to define the interfacial region. We used three residues with the highest energy to represent the patch. The mean energy of the three highest-energy residues in a patch was calculated and used for patch analysis. The number of total patches was also decreased, since some surface patches, which have the same three highest-energy residues, were considered to be one patch. As a result, the three highest-energy residues of the top energy patch belong to interfacial residues in four small protomers (OMTKY3/elastase, OMTKY3/chymotrypsin, streptomyces inh. and hen egg lysozyme). Two of the three highest-energy residues of the top energy patch are interfacial residues in the other four protomers (Fasciculin-II, Barstar, BPTI and Chymotrypsin inh.2). For the large protomers, the situation is different. The highest-energy residue on the protein surface is located in the non-interfacial region for every protomer. One of the three highest-energy residues of the top energy patch is the interfacial residue in four large protomers (Elastase, Barnase, Trypsin, and Subtilisin BPN′). None of the three highest-energy residues of the top energy patch are interfacial residues for the other four protomers (α-chymotrypsin, Acetylcholinesterase, Fab D44.1 and Subtilisin novo). Thus the interfacial regions of the small protomers could be effectively predicted by only considering residues with the highest energy in the surface patches. However, this method would not work for the large protomers.

Prediction of Hot Spot Residue by Searching for Residue with Highest Energy

At the protein–protein interface, only a few residues typically contribute the majority of the binding affinity.[23–25] The energetically important residues were designated the 'hot spot.' Some hot spot residues are quite solvent accessible, while others are not (Table III). The relatively buried hot spot residues, such as L92W in antibody D1.3, may be important in maintaining the structure of the interface so that other interface residues can form favorable interactions with the partner protein.

TABLE III. Analysis of Side-Chain Energy Scores for Key Interaction Residues

Protein	Partner Protein	Size of Interface[a]	Hot Spot Residues[b]	Relative Accessibility	Rank[c]
Barnase	Barstar	20	102H	62%	6
			87R	5.9%	13
			83R	27%	11
			27K	46%	12
			59R	84%	1
Barstar	Barnase	18	39D	79%	7
			35D	88%	3
			29Y	79%	1
Antibody D1.3[d]	Lysozyme	16	H101Y	64%	5
			L92W	47%	14
			H100D	69%	9
Hen egg lysozyme	D1.3	19	121Q	86%	1
BPTI	trypsin	12	15K	96%	1

[a]The interface was composed of residues with decreased side-chain accessible surface area upon association even if they were not defined as surface residues.
[b]The changes in experimental binding energy on mutation to alanine were >2.5 kcal/mol. Alanine scanning mutagenesis and binding data were obtained from ref. 23 for barnase/barstar, ref. 25 for lysozyme/Antibody D1.3, and ref. 24 for BPTI respectively.
[c]The rank of the hot spot residue among interface residues in terms of the calculated side-chain energy score.
[d]The PDB code of antibody D1.3 is 1a7r; the code of D1.3/lysozyme complex is 1vfb.

We calculated these residues to have low side-chain energy. The exposed hot spot residues were often calculated to have the highest energy among interface residues. In fact, the highest-energy residue at the interface is a hot spot residue in four out of five analyzed proteins. Thus analysis of the side-chain energy scores of interface residues could be helpful in predicting which residues are hot spot residues. The residue with the highest energy at the interface is highly likely to be the hot spot residue. Even with no knowledge of the protein interface, we could identify possible hot spot residues by searching for the residues with the highest energy score on the protein surface, which is usually located at the interface for the small protomers.

Decomposing the Scoring Function

It is of interest to discover which energy terms in the scoring function contribute to the high energy of the observed interfaces. Originally, our scoring was designed to calculate the energy of a rotamer placed on the modeled position. Side-chain energy of the modeled residue was derived from all of the rotamers of the residue type in a nonlinear way, as described by eq. (3), to account for the entropy effect. Eq. (3) is consistent with eq. (2), which was minimized to balance the weights of the energy terms. When the rotamer was evaluated with any function other than the trained scoring function, eq. (3) could no longer be used to derive the side-chain energy. Instead, the rotamer, which had the smallest root mean square difference (RMSD) value with respect to the native side chain conformation, was evaluated using a particular energy term of the scoring function. The calculated value was considered to be the side-chain energy score and was used for patch analysis (Table IV). The term corresponding to disulfide bridges, which are not commonly found on the protein surfaces, was not considered. The interface patches were calculated as high energy (the mean value of the decile ranks of the 16 monomers was <5) by four energy terms out of ten: the reference value of amino acids (ΔG_{ref}), the observed frequency of the residue type and the rotamer given the backbone conformation ($f_1 \times f_2$), the fraction of buried surface of non-hydrogen-bonded hydrophilic atoms (F_{phi}) and the hydrogen bond energy (E_{hbond}). The term F_{phi} makes a positive contribution to the high ranks of the interface patches, which indicates that the interface regions are rich in buried or partially buried non-hydrogen-bonded hydrophilic atoms. Tight binding affinity could be achieved when these hydrophilic atoms form hydrogen bonds with the partner protein. The positive contribution of the $f_1 \times f_2$ term reveals that interface residues have a higher intrinsic energy than other surface residues. However, when the four energy terms, F_{phi}, E_{hbond}, $f_1 \times f_2$ and ΔG_{ref} were used together, we obtained only a slightly higher mean rank for the 16 protomers (1.9) than we obtained when using the full energy terms of the scoring function (2.3). Again, we used the energy of the rotamer with the smallest RMSD value in the patch analysis. In the standard procedure, the side chain energy score was derived from all of the rotamers of the modeled residue to account for the entropy effect, and the mean rank of the interface patches was found to be 1.9 as well. Obviously, the entropy effect also contributes to the high energy of the observed interfaces. As indicated by Cole and Warwicker,[15] interfacial regions are less flexible than the rest of the protein surface. When the entropy effect was considered, the free energy of the non-interfacial residues was found to decline more significantly than that of the interfacial residues.

The effects of the different energy terms were correlated. For example, the observed interfaces were predicted to have slightly lower energy than other surface patches, as evaluated individually by the contact surface or overlap volume (Table IV). However, the observed interfaces were predicted to have high energy (mean rank 4.2) when the

TABLE IV. Ranks of Observed Interfaces Evaluated by Individual Energy Term of Scoring Function

Unbound Monomers	$S_{contact}$	$V_{overlap}$	E_{hbond}	E_{elec}	ΔS_{pho}	ΔS_{phi}	$\Delta(F_{phi})^{30}$	$V_{exclusion}$	$\ln(f_1 \times f_2)$	ΔG_{ref}	RA[a]
Elastase	6	4	2	9	8	8	1	5	2	2	8
OMTKY3	1	10	3	9	1	5	2	4	2	5	1
α-Chymotrypsin	10	1	9	9	9	3	2	3	1	1	8
OMTKY3	1	10	1	9	1	3	1	6	1	6	1
Acetylcholinesterase	7	2	3	1	8	8	2	8	4	1	5
Fasciculin-II	7	4	8	6	5	2	1	4	2	1	1
Barnase	10	2	7	10	10	2	4	2	1	1	5
Barstar	8	2	2	1	6	9	6	10	6	4	5
Subtilisin BPN′	4	7	1	3	8	8	1	10	2	4	7
Streptomyces inh.	3	6	1	6	3	8	3	3	1	1	1
Trypsin	9	1	3	6	8	5	1	6	4	3	10
BPTI[b]	1	9	3	3	1	10	3	8	5	7	1
Subtilisin novo	3	9	2	4	5	10	2	9	2	6	4
Chymotrypsin inh.2	2	10	1	3	3	9	5	8	3	2	1
Fab D44.1	10	1	7	3	8	1	2	2	2	3	9
Hen egg lysozyme	1	9	2	6	1	8	1	3	7	8	1
Mean	5.2	5.4	3.4	5.5	5.3	6.2	2.3	5.7	2.8	3.4	4.3

[a]Relative accessibility (RA) was used as the evaluation method. The definition of each energy term was the same as used in eq. (1). The rotamer, which had the smallest RMSD value with the native side chain conformation, was evaluated by a particular energy term of the scoring function. The calculated value was considered to be the side-chain energy score and was used for patch analysis. The ranks of the interface patches relative to other surface patches were divided into ten equally sized categories.

two energy terms were used together, which indicates that the interfacial regions are packed more poorly than the other surface regions. Additionally, the effect of a single energy term is dependent on different proteins. For example, the interface of barstar is rich in negatively charged residues, which could form favorable electrostatic interactions with the positively charged residues clustered at the interface of banase. Electrostatic interaction energy (E_{elec}) is significantly higher in the interfacial region of barstar than in other surface regions. For other protomers, on average, we obtained low electrostatic interaction energy in the interfacial regions.

Effect of Amino Acid Composition

The observed interfaces were evaluated as having high rank by the single energy term ΔG_{ref}, which indicates that the reference values of the interfacial residues are lower than those of other surface residues (ΔG_{ref} is negative in the scoring function). In our previous study,[1] the reference values of the amino acids were derived from the optimization procedure. Hydrophobic residues and large residues such as arginine were shown to have low reference values. These facts are consistent with the composition of the interface regions. In the 16 protomers, 32% of the interface residues and 26% of the non-interface residues are hydrophobic residues (Ile, Leu, Cys, Pro, Trp, Val, Phe, Val, and Met). Arg constitutes 6.3% of the interface residues and 4.3% of the non-interface residues. In general, charged residues (Asp, Glu, Lys and Arg) are not as common in the interface regions (19%) as they are in other surface regions (21%).

However, the high energy of the observed interfaces is irrelevant to the amino acid composition in the interface regions. Although hydrophobic residues usually have a lower reference value than hydrophilic residues, hydrophobic interactions are predicted to be favorable in the scoring function (the weight term is negative). As a result, the effects of the two energy items cancel each other to a large degree. In fact, all of the 20 amino acids have a higher mean energy score in the interface regions than in other surface regions (Table V). When the mean energy of each amino acid on the protein surface is used as the energy of the corresponding interface residues and the energy of other surface residues is calculated as usual, we obtain the mean rank of 4.8 for the observed interfaces. We could obtain a higher mean rank (1.6) by removing the terms S_{pho} and $V_{exclusion}$ from the scoring function used in the standard procedure (1.9). However, the calculation results would no longer be independent for the amino acid composition of the interfacial regions. The mean rank is unchanged by removing the S_{phi} term, which has a small weight in the scoring function. However, we obtained a lower mean rank of 2.5 when the terms S_{pho}, $V_{exclusion}$, S_{phi} and ΔG_{ref} were removed altogether.

Effect of B-factors on Energy of Interface Patches in Unbound and Bound Monomers

Since other groups have completed similar calculations on the monomers extracted from the complexes,[9,15] we decided to compare the calculated results for the unbound and bound monomers (Table VI). We expected the interfaces of the bound monomers to have a higher rank than those of the unbound monomers since the interface residues in the complex structures were adjusted to optimize the interactions between the two molecules. Some favorable interactions in the interfacial regions of the unbound monomers could be broken during molecular association. To our surprise, the observed interfaces of bound monomers have an even lower decile rank than those of the unbound monomers. Remarkably, the interface patch of the unbound antibody Fab D44.1 was ranked at 2 com-

TABLE V. Comparison Between Mean Energy of Interface Residues and that of Other Surface Residues for Each Amino Acid

Amino Acids	Interface		Non-Interface	
	No. of Residues	Energy (kcal/mol)	No. of Residues	Energy (kcal/mol)
Ala	10	−0.23	124	−0.36
Arg	19	−0.35	78	−1.01
Asn	25	−0.62	136	−0.74
Asp	12	−0.50	91	−0.84
Cys	11	−2.25	29	−2.88
Gln	8	0.07	80	−0.69
Glu	16	0.01	87	−0.52
Gly	30	−0.10	173	−0.33
Ile	10	−1.16	50	−2.44
Leu	17	−1.88	97	−2.29
Lys	10	−0.66	127	−0.99
Met	10	1.41	18	−1.22
Phe	13	−1.64	38	−2.58
Pro	14	0.72	100	−0.46
Trp	9	−1.37	25	−2.22
Val	12	−1.76	117	−1.87
Ser	24	−0.05	234	−0.07
Thr	17	−0.34	136	−0.82
Tyr	24	−0.76	56	−2.42
His	10	−0.60	27	−1.22

The unbound monomers of the eight protein complexes were used for the statistical analysis.

TABLE VI. Comparison of Ranks of Observed Interfaces Relative to Other Surface Patches Between Unbound and Bound Monomers

Proteins	Unbound	Bound
Elastase	1	2
OMTKY3	1	1
α-Chymotrypsin	3	6
OMTKY3	1	1
Acetylcholinesterase	2	1
Fasciculin-II	1	1
Barnase	4	5
Barstar	1	2
Subtilisin BPN′	3	5
Streptomyces inh.	1	1
Trypsin	2	1
BPTI	1	1
Subtilisin novo	6	4
Chymotrypsin inh.2	1	1
Fab D44.1	2	10
Hen egg lysozyme	1	2
Mean	1.94	2.76

The ranks were calculated on a scale of 1 to 10.

pared to a rank of 10 for the monomer extracted from the antibody–lysozyme complex. The mean rank of the observed interfaces of the 16 bound monomers is lower (2.8) than that of the unbound monomers (1.9).

By decomposing the scoring function, we found that hydrophilic atoms at the interfaces of the bound monomers are more solvent accessible than those of the unbound monomers. The interfacial regions of the bound monomers were also calculated to have higher energy than those of unbound monomers by the single energy term of electrostatic interaction energy. These facts indicate that the interface residues are indeed adjusted to optimize intermolecular interactions in the complex structures by sacrificing intramolecular interactions. However, the interfacial regions of the bound monomers were also calculated to have low energy by a single energy term, such as atom overlap volume, residue intrinsic energy or hydrogen bond energy. This is because the interface residues in the bound monomers are buried in the complex structures and have low B-factors. At the same time, the interface residues in the unbound monomers are solvent accessible and have high B-factors. Residues with high B-factors adopt unusual conformations more frequently and have more atomic clashes than those with low B-factors.[26] These residues were calculated to have high energy by the single energy term of atom overlap volume or residue intrinsic energy. Especially, the value for hydrogen bond energy was found to be more sensitive to small atomic coordinate errors than that of electrostatic interaction energy, which accounts for the different results as calculated using the two energy terms individually.

In the unbound structure of antibody Fab D44.1, the OE1 atom of the interface residue B35Glu overlaps the NE atom of B47Trp. However, the distance between the two hydrogen-bonded atoms increases from 2.45 to 2.82 Å in the complex structure. Additionally, the interface residue B100Asp adopts a very rare conformation in the unbound structure, while it adopts a typical conformation in the complex structure. In fact, the high energy of the interface patch of unbound Fab D44.1 is mostly due to two residues, B35Glu and B100Asp. When side chain energy was calculated using a single energy term, the overlap volume (V_{overlap}) or residue intrinsic energy ($f_1 \times f_2$), the rank of the Fab D44.1 interface patch was found to be 1 or 2 respectively for the unbound monomer. The ranks were 3 and 5 for the bound monomer. Since the interface residues of the unbound monomers are supposed to have B-factors similar to those of other surface residues, the results calculated from the unbound monomers are more reliable than those calculated from the bound structures.

Discussion

We have shown that residues at the interfaces of protein–protein complexes have higher energy scores than other surface residues. However, no correlation between this property and the binding free energy was observed. We also tried calculating the energy score for both side-chain and backbone atoms for each surface residue. But the ranked ordering of the observed interfaces relative to other surface patches was similar to the results obtained when evaluating only side-chain atoms. Although the calculation results could be affected by the inaccuracy of the scoring function, patch analysis of side-chain energy on the protein surface still could be helpful in identifying the possible protein–protein interface. The procedure is valuable because it is particularly powerful for small protomers with spherical shapes, while other computational

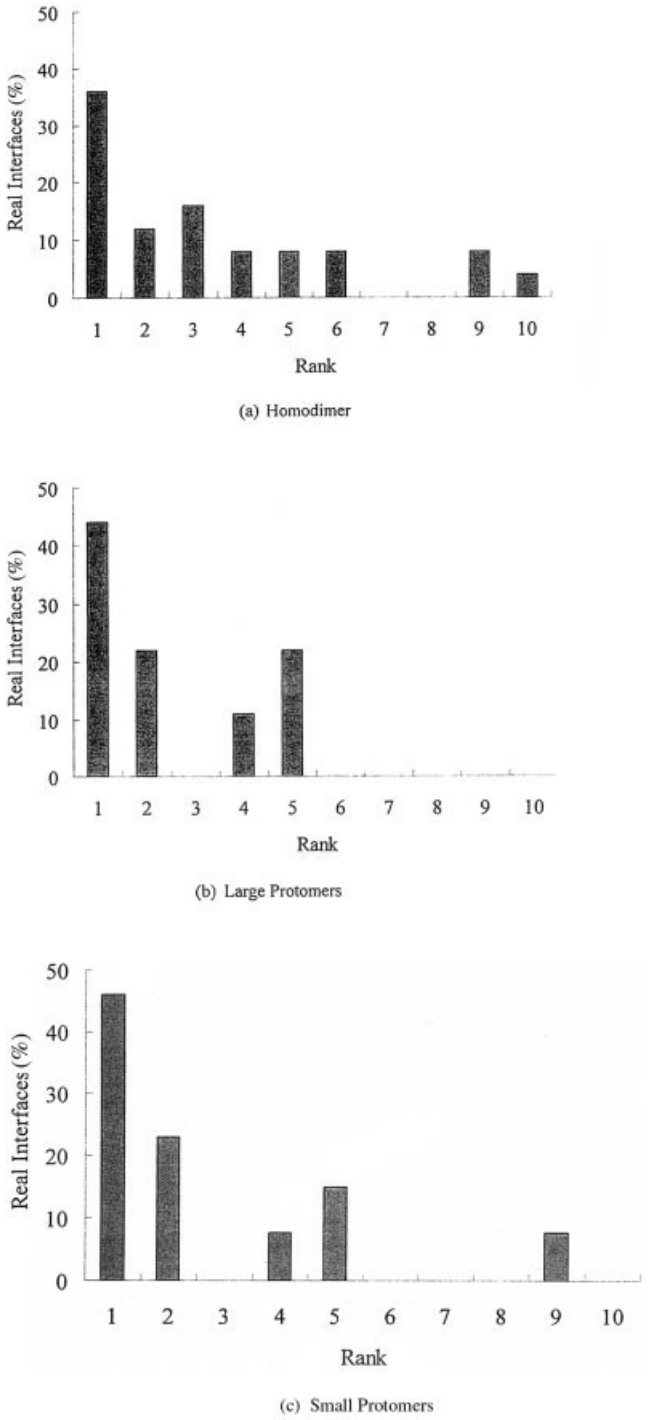

Fig. 3. Distributions of the ranked ordering of the observed interface patches relative to other surface patches for homodimers and hetero-complexes. (**a**) homo-dimers, (**b**) large protomers, (**c**) small promoters. The 9 large and 13 small protomers, which had a sequence identity of <35%, were extracted from 14 hetero-complexes and 25 homo-dimers were obtained from ref. 15.

tools encounter difficulty in determining the functional site on the convex protein surface. The interfaces of the large protomers can be identified easily since they are usually located at the main concave surface. The ligand binding sites of enzymes could be accurately predicted by placing molecular probes (small organic molecules) on a protein surface and finding regions with the most favorable binding free energy,[27–29] or by simply searching for pockets and cavities at the protein surface.[30–32]

Jones and Thornton have characterized protein–protein interaction sites in complexes of known structure to evaluate what differentiates them from other sites on the protein surface.[9] They calculated six parameters, solvation potential, residue interface propensity, hydrophobicity, planarity, protrusion and accessible surface area for the observed interfaces and other surface patches defined on the monomers from a dataset of homo-dimers and hetero-complexes. Using essentially the same datasets, Cole and Warwicker have found that interfacial regions are less flexible than the rest of the protein surface.[15] Here, based on side-chain energy scores, the ranked ordering of the interface patches relative to other surface patches was calculated for the 25 homodimers and 14 hetero-complexes described by Cole and Warwicker. The mean values of the decile ranks are 3.4, 2.4, and 2.7 for the homo-dimers, large protomers and small protomers respectively (Fig. 3). For the same datasets, the mean ranks calculated by Cole and Warwicker are 2.0, 2.0 and 4.0 respectively. Jones and Thornton obtained higher ranks than those obtained in this study for the interfaces of the homo-dimers by analysis of residue interface propensity or planarity of the surface patches. They obtained similar ranks as those obtained in this study for the interfaces of the small protomers by calculating protrusion index or residue accessible surface area. However, here the interfaces of the large protomers were ranked in higher order by calculating side-chain energy scores than by calculating any of the six parameters used by Jones and Thornton.

In general, the interfaces of the homo-dimers are ranked lower by our algorithm than in other studies, while those of the hetero-complexes are ranked at higher or similar values. We found that homo-dimers usually only exist in the complexed state, while the monomers of hetero-complexes can also exist independently. The interfaces of the homo-dimers are larger and more hydrophobic than those of the hetero-complexes. The tight binding affinity can be easily achieved by accumulating favorable hydrophobic interactions. As a solvable protein, the monomers of the hetero-complexes cannot have a completely hydrophobic interface. Interfacial residues have high side-chain energy in the unbound monomers, which may be a strategy used by hetero-complexes to ensure that the free energy of the two interacted proteins declines significantly upon association. In addition, other groups analyzed the monomers extracted from the protein complexes.[9,15] We also used the extracted monomers here. Interface residues were found to be buried in the complex structures. Other surface residues were more likely to have high energy than the interface residues due to atom coordinate errors, a phenomenon which was exacerbated because low quality structures (resolution >2.5 Å) were included in the datasets. Some surface residues were invisible in the crystal structures. For example, in the crystal structure of trpysin/Bowman-Birk inhibitor (PDB code 1tab), only the

trypsin binding domain of the inhibitor (36 of 82 residues) was included in the model. The residues, which interacted with the missing domain, could be calculated to have high energy. In addition, the interface residues have much lower B-factors than other surface residues. As a result, the interface patch was ranked 9. Excluding Bowman-Birk inhibitor, we obtained a mean rank of 2.2 for the small protomers. Our criteria for selecting surface and interface residues were similar to those used by Jones and Thornton but different from those used by Cole and Warwicker. Cole and Warwicker classed surface residues as those with an accessible surface area (ASA) in the monomer ≥ 0.1 Å^2. The interface was defined as the set of residues for which the ASA decreased by ≥ 0.1 Å^2 upon association. Solvent vectors[9] were not used to define surface patches, which may have resulted in sampling patches through the center of a protein. In such cases, many patches composed of residues with low accessibility were generated. If we used these criteria, the interface patches would be ranked even higher than in the standard procedure since our scoring function evaluated the buried residues as low energy. The mean ranks were 2.2, 1.4 and 1.7 for the interfaces of homo-dimers, large protomers and small protomers respectively.

CONCLUSIONS

Side chains at the interfaces of protein–protein complexes were found to have higher energy than other surface residues. The higher side-chain energy of interface residues calculated for the unbound monomers may be a strategy used by hetero-complexes to ensure that the free energy of the two interacting proteins declines significantly upon association. The calculation results are independent of the amino acid composition in the interface regions. Each of the 20 amino acids has a higher energy score in the interface regions than in other surface regions. Interface residues in the monomers extracted from complex structures have lower B-factors than other surface residues, and the side-chain energy can be underestimated. To make a fair comparison, unbound monomers should be used in the calculation. Patch analysis of side-chain energy on the surface of an isolated protein structure can be useful in identifying the protein–protein interface, especially for small protomers, in situations for which the structure of the partner protein is not known. In addition, the energy scores of residues in the low quality proteins can be overvalued due to atom coordinate errors or missing nearby residues. Other information besides side-chain energy scores should be considered when deriving the putative protein biding sites.

REFERENCES

1. Liang S, Grishin NV. Effective scoring function for protein sequence design. Proteins 2004;54(2):271–81.
2. Camacho CJ, Vajda S. Protein docking along smooth association pathways. Proc Natl Acad Sci USA 2001;98(19):10636–41.
3. Smith GR, Sternberg MJ. Prediction of protein-protein interactions by docking methods. Curr Opin Struct Biol 2002;12(1):28–35.
4. 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 USA 1992;89(6):2195–9.
5. Moont G, Gabb HA, Sternberg MJ. Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins 1999;35(3):364–73.
6. Palma PN, Krippahl L, Wampler JE, Moura JJ. BiGGER: a new (soft) docking algorithm for predicting protein interactions. Proteins 2000;39(4):372–84.
7. Norel R, Sheinerman F, Petrey D, Honig B. Electrostatic contributions to protein-protein interactions: fast energetic filters for docking and their physical basis. Protein Sci 2001;10(11):2147–61.
8. Kimura SR, Brower RC, Vajda S, Camacho CJ. Dynamical view of the positions of key side chains in protein-protein recognition. Biophys J 2001;80(2):635–42.
9. Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches. J Mol Biol 1997;272(1):121–32.
10. Jones S, Thornton JM. Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 1997;272(1):133–43.
11. Young L, Jernigan RL, Covell DG. A role for surface hydrophobicity in protein-protein recognition. Protein Sci 1994;3(5):717–29.
12. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. Protein Sci 1997;6(1):53–64.
13. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci USA 1996;93(1):13–20.
14. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285(5):2177–98.
15. Cole C, Warwicker J. Side-chain conformational entropy at protein-protein interfaces. Protein Sci 2002;11(12):2860–70.
16. Fernandez A, Scheraga HA. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. Proc Natl Acad Sci USA 2003;100(1):113–8.
17. Xu D, Tsai CJ, Nussinov R. Hydrogen bonds and salt bridges across protein-protein interfaces. Protein Eng 1997;10(9):999–1012.
18. Ringe D. What makes a binding site a binding site? Curr Opin Struct Biol 1995;5(6):825–9.
19. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235–242.
20. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 1999;285(4):1735–47.
21. Dunbrack RL, Jr., Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6(8):1661–81.
22. Braden BC, Souchon H, Eisele JL, Bentley GA, Bhat TN, Navaza J, Poljak RJ. Three-dimensional structures of the free and the antigen-complexed Fab from monoclonal anti-lysozyme antibody D44.1. J Mol Biol 1994;243(4):767–81.
23. Schreiber G, Fersht AR. Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. J Mol Biol 1995;248(2):478–86.
24. Castro MJ, Anderson S. Alanine point-mutations in the reactive region of bovine pancreatic trypsin inhibitor: effects on the kinetics and thermodynamics of binding to beta-trypsin and alpha-chymotrypsin. Biochemistry 1996;35(35):11435–46.
25. Dall'Acqua W, Goldman ER, Lin W, Teng C, Tsuchiya D, Li H, Ysern X, Braden BC, Li Y, Smith-Gill SJ and others. A mutational analysis of binding interactions in an antigen-antibody protein-protein complex. Biochemistry 1998;37(22):7981–91.
26. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins 2000;40(3):389–408.
27. Dennis S, Kortvelyesi T, Vajda S. Computational mapping identifies the binding sites of organic solvents on proteins. Proc Natl Acad Sci USA 2002;99(7):4290–5.
28. Kortvelyesi T, Dennis S, Silberstein M, Brown L, 3rd, Vajda S. Algorithms for computational solvent mapping of proteins. Proteins 2003;51(3):340–51.
29. Silberstein M, Dennis S, Brown L, Kortvelyesi T, Clodfelter K, Vajda S. Identification of substrate binding sites in enzymes by computational solvent mapping. J Mol Biol 2003;332(5):1095–113.
30. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci 1998;7(9):1884–97.
31. Brady GP, Jr., Stouten PF. Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Des 2000;14(4):383–401.
32. Ondrechen MJ, Clifton JG, Ringe D. THEMATICS: a simple computational predictor of enzyme function from structure. Proc Natl Acad Sci USA 2001;98(22):12473–8.