J Mol Model (2004) 10:44–54
DOI 10.1007/s00894-003-0168-3

ORIGINAL PAPER

Ying Gao · Renxiao Wang · Luhua Lai

Structure-based method for analyzing protein–protein interfaces

Received: 2 June 2003 / Accepted: 15 October 2003 / Published online: 22 November 2003

Abstract Hydrogen bond, hydrophobic and vdW interactions are the three major non-covalent interactions at protein–protein interfaces. We have developed a method that uses only these properties to describe interactions between proteins, which can qualitatively estimate the individual contribution of each interfacial residue to the binding and gives the results in a graphic display way. This method has been applied to analyze alanine mutation data at protein–protein interfaces. A dataset containing 13 protein–protein complexes with 250 alanine mutations of interfacial residues has been tested. For the 75 hot-spot residues ($\Delta\Delta G \geq 1.5$ kcal mol^{-1}), 66 can be predicted correctly with a success rate of 88%. In order to test the tolerance of this method to conformational changes upon binding, we utilize a set of 26 complexes with one or both of their components available in the unbound form. The difference of key residues exported by the program is 11% between the results using complexed proteins and those from unbound ones. As this method gives the characteristics of the binding partner for a particular protein, in-depth studies on protein–protein recognition can be carried out. Furthermore, this method can be used to compare the difference between protein–protein interactions and look for correlated mutation.

Y. Gao · R. Wang · L. Lai
State Key Laboratory of Structural Chemistry
for Stable and Unstable Species,
College of Chemistry and Molecular Engineering,
Peking University,
100871 Beijing, China

L. Lai (✉)
Center for Theoretical Biology,
Peking University,
100871 Beijing, China
e-mail: lhlai@pku.edu.cn
Tel.: +86-10-62757486
Fax: +86-10-62751725

Present address:
R. Wang, Medical Chemistry and Comprehensive Cancer Center,
University of Michigan,
1500 E. Medical Center Drive Ann Arbor, MI 48109-0934, USA

Keywords Protein–protein interaction · Interface analysis · Hot spot · Correlated mutation · PP_SITE

Introduction

Biomolecular recognition and protein-interaction networks are key issues in understanding cellular functions. Elucidating the structure, interactions and functions of all proteins within cells and organisms is an ambitious goal of proteomics. Finding the functional sites in proteins is an important step in reaching this goal. After finding the functional sites on a protein surface, identifying individual residues that dominate function is the key step for detailed analysis. Structural and evolutionary information is the foundation of such work. [1]

Many methods have been developed for looking for functional sites. Some of them use only sequence information. For example, Bock and Gough (2001) [2] used a Support Vector Machine learning system trained with sequence and associated physicochemical properties to find functional sites. Additionally, Kini and Evans (1995), [3] Casari et al. (1995), [4] Pazos et al. (1997) [5] and Gallet et al. (2000) [6] also reported their sequence-based methods to search for functional epitopes in proteins. Sequence is the starting point of such analyses, but they are limited for two reasons. Firstly, many functions involve large interfacial areas, rather than short local sequence motifs. Secondly, functional analogies can be specious, especially when sequence identity falls below 40%. Thus, addition of structural information will generally give better results. A good example is the Evolutionary Trace (ET) approach combined with structural information. ET ranks the residues in a protein sequence by evolutionary importance with phylogenetic information. It then maps those residues ranked at the top onto a representative structure. If these residues form structural clusters, they can be identified as functional determinants. The work of Madabushi et al. (2002) [7] is a good example. Aloy and Russell (2003) [8, 9] have built up a web-based method, InterPreTS, to predict protein–

protein interaction through tertiary structure. Methods developed by Zhou et al. (2001), [10] Fariselli et al. (2002) [11] and Pupko et al. (2002) [12] all belong to this class.

Therefore, the function of these methods is to find a linear stretche of sequence or structural clusters that are important for protein function. In other words, the methods mentioned above are mainly to look for the part on the surface of proteins that interacts with another protein. Hereinafter, we will introduce methods that can find individual residues dominant to function. A known structure is a prerequisite. As mutational analysis is a major experimental method for finding functional sites at protein–protein interfaces, all these methods have compared their results with mutational experiments.

By using alanine-scanning mutagenesis [13] to probe the energetic contribution of individual side chains to protein binding, Clackson and Wells [14] found that, despite of the large size of the binding interface, single residues could still contribute a large fraction of the binding free energy. Bogan and Thorn (1998) [15]also found that there were "hot spots" (i.e. amino acids whose replacement by alanine is unfavorable) of binding energy made up of a small subset of residues in the dimeric interface according to a database [16] of 2,325 alanine mutants in 22 protein–protein complexes. A systematic analysis [17, 18] of a wide range of protein–protein interfaces has shown a diversity of interaction patterns and no general rules for hydrophobicity, polarity, or shape, which can be used as a basis for predicting which atoms will participate in hot spots. Since reliable prediction of key residues in the interface has immediate applications in both rational design of therapeutic agents and protein engineering, considerable effort has been invested over the past few years in schemes designed to identify hot spots on protein surfaces.

A number of such methods have been developed, which fall generally into three classes. The first class includes methods that estimate the free energy of association directly or changes in the binding free energies as a result of mutating the residues of the interacting molecules. Computational alanine scanning reported by Massova and Kollman [19, 20, 21] is a good example. They studied the interaction free energies in protein–protein complexes from a single simulation and estimated of the individual contribution of each residue to binding. First, conventional MD simulations were performed to generate a representative sample of conformations of the original complex. Then a combined Poisson–Boltzmann and solvent accessible surface area (PB/SA) solvation model was applied to calculate the solvation energies of the complex and the separated proteins, and a molecular mechanics force field was used to calculate other energies. The entire process was then repeated for the alanine mutants, with their structures being obtained by simply removing the appropriate side-chain atoms, leaving only the methyl groups. Verkhivker and Bouzida et al. (2002) [22] used energy landscape analysis to study structural and energetic aspects of molecular recognition. The distinct difference between this and Massova's analysis is that the former used the simplified energy function in conjunction with Monte Carlo simulations to sample the conformational space and the resulting conformational states were evaluated with a detailed binding free energy model, which includes the molecular mechanics AMBER force field and the solvation energy term based on continuum generalized Born and solvent accessible surface area (GB/SA) solvation model. The two methods mentioned above gave good qualitative results in those systems to which they were applied, while Massova's analysis was designed to predict mutation quantitatively. [21] However, the high computational cost of such a thorough approach and the difficulty in operation (parameter selection, data processing, etc.) make it unsuitable for users not familiar with this field.

The second class encompasses approaches that use the sequence as the starting point of the analysis.

The first step is to predict protein structure and functional epitopes using various sequence comparisons if no structural information is known, as described above. When structural information is known, sequence and structure comparison can be used to identify the conserved residues, especially polar residues, as energetic hot spots at the intermolecular interface. Hu et al. (2000) [23] have selected a structural non-redundant subset of 11 interface families with 97 protein–protein interfaces. Each interface of a family is superimposed structurally (C_α atoms only) on the interface that represents the family. The results show that all families have their own set of conserved residues. Hu's method can be regarded as a qualitative method for finding hot spots at protein–protein interfaces. As this kind of method requires sequence and structure alignment with a known database, it is of statistical significance, but may be difficult to apply to the analysis of a single protein–protein complex.

The third class includes methods that make grids around the binding interface and use probes to explore the properties of protein. In general, these methods use knowledge-based simplified models to evaluate binding, such as hydrogen bonding property or hydrophobicity. They can also be expanded to the whole surface of the protein. Based upon surface hydrophobicity, Young et al. (1994) [24] and Villoutreix et al. (1998 [25], 2001 [26]), used a simplified protein model, consisting of only the Cα coordinates, to represent the geometry of each residue. Each residue type was assigned a hydrophobic value. For each lattice position exterior to the molecular surface, the relative strength, for which a ligand might bind at that site, was the sum of hydrophobic values of residues within 7.5 Å of that position. Strictly speaking, this method is not a way to find hot-spot residues, but an enumeration of key functional regions of protein. They carried out enumeration of binding sites for some systems such as the model structures of C4b-binding protein (C4BP) [25] and Protein C/activated Protein C [26], and identified several binding sites that have already been established experimentally. Making grids in binding sites is a general practice in studying the interaction between

proteins and small molecules, for example, the well-known program GRID [27].

From the above analysis, we can see that no general algorithm has yet been developed that can predict hot spots based solely on their shape or composition and be easily applied. We have developed a method called PP_SITE for protein–protein interface analysis, which was briefly reported in the previous paper [28]. Here, we will give the details of the method and its application in alanine scanning analysis at protein–protein interface. PP_SITE only uses hydrogen bonds and hydrophobic characteristics to pick out key residues at protein–protein interfaces and decompose the contribution of atoms in hot-spot residues, which can describe the properties of the protein–protein interface and be visualized easily with graphics software.

Materials and methods

Dataset of protein–protein complexes

Mutating the amino acid of interest using the technique of site-directed mutagenesis can reveal the roles played by individual amino acid side chains in determining the strength of a protein–protein complex. Alanine scanning is particularly important. In this technique, the amino acids are mutated, one by one, to alanine, the side chain of which consists only of a single methyl group. Comparing the binding affinity of the real (wild-type) protein with the alanine mutant protein then gives an indication of importance of that amino acid's side-chain's interactions for the binding affinity if removal of the side chain does not cause any drastic structural rearrangement of the complex. Data from alanine-scanning mutagenesis are now accessible through the Internet. ASEdb is a database of alanine mutations collected by Thorn and Bogan (2001) [16]. Though the total number of mutations was large, the number of mutations on protein–protein interfaces whose structures have been determined was small. From ASEdb, we selected 13 protein–protein complexes that had mutations on interfaces. Their PDB codes with the chains concerned are 1a4y(A:B), 1ahw(AB:C), 1brs(A:D), 1bxi(A:B), 1cbw(ABC:D), 1dan(LH:TU), 1dfj(E:I), 1dvf(AB:CD), 1gc1(C:G), 1jck(AC:BD), 1vfb(AB:C), 3hfm(LH:Y), 3hhr(A:BC). We compared the similarity of the complexes in this dataset using the program ALIGN [29]. 1vfb is the Fv fragment of mouse monoclonal antibody D1.3 complexed with hen egg lysozyme. 3hfm is the HyHEL-10 Fab–hen egg lysozyme complex. Chain AB of 1vfb has 30% sequence similarity with chain LH of 3hfm, but lysozyme uses different sites to interact with these two antibodies and the mutations in two antibodies do not coincide. 1a4y is a ribonuclease inhibitor–angiogenin complex and 1dfj is a ribonuclease inhibitor complexed with ribonuclease A. Chain B of 1a4y has 31% sequence similarity with chain E of 1dfj and chain B of 1a4y has 77% sequence similarity with chain I of 1dfj. They have similar structures, but the mutated residues in two complexes are not repeated. Thus, we can say that this dataset is non-redundant. We used our program and Δ ASA (change of area of solvent accessible in the binding) to judge if a residue was at the interface. There are 391 mutations, with 250 residues at interfaces. The difference in free energy of binding between the wild-type (wt) and mutant protein ($\Delta\Delta G = \Delta G_{mut} - \Delta G_{wt}$) is most important. For 141 mutations not at an interface, the distribution of $\Delta\Delta G$ is from −0.8 to 3.75 kcal mol^{-1}, 0.20 ± 0.51 kcal mol^{-1}, in which 22 mutations whose $\Delta\Delta G \geq 0.5$ kcal mol^{-1}. For 250 mutations at interfaces, the distribution of $\Delta\Delta G$ is from −0.9 to 7.7 kcal mol^{-1}, 1.20±1.47 kcal mol^{-1}, in which 104 mutations whose $\Delta\Delta G < 0.5$ kcal mol^{-1} and 75 whose $\Delta\Delta G \geq 1.5$ kcal mol^{-1}. According to these statistics and [30], here we define hot-spot residues with $\Delta\Delta G \geq 1.5$ kcal mol^{-1}, and warm residues with $\Delta\Delta G =$ 0.5–1.5 kcal mol^{-1}. We will use PP_SITE to predict hot-spot residues and warm residues in these 13 complexes and compare with mutation data. (The results can be download from the ftp site ftp://mdl.ipc.pku.edu.cn/pub/software/pp_site/ala_result.doc.)

In order to test the tolerance to conformational changes upon binding simply, we compare results from complexed and unbound proteins. Here we utilize a set of 31 complexes with one or both of their components available in an unbound form. This dataset comes from [31]. Firstly, the protein whose code is 1jel had been superseded by 2jel and 2ssi has been superseded by 3ssi. Secondly, we compared the similarity among these complexes with ALIGN. If two unbound proteins have sequence similarity more than 25% and use the same sites to interact with their partners, we will get rid of the one with lower resolution. Finally, in order to complement this dataset, we searched PDB to find those unbound proteins that are not in this list. Currently, this dataset contains 26 complexes. Fifteen complexes are enzyme-inhibitors, five are antibody-antigens, and the remaining six are of other types (see Appendix). We will use our program PP_SITE to deduce key residues of proteins in this dataset. Key residues include hot-spot residues and warm residues as defined above.

PP_SITE

PP_SITE is developed based on POCKET, one module of a multi-purpose program LigBuilder [28, 32] for structure-based drug design, which was originally designed for analyzing the interactions between proteins and small molecules.

The main function of PP_SITE is to analyze the binding interface of proteins and deduce hot spots—key residues in the binding interface. As we know, the energetics of protein–protein interaction arise from favorable intermolecular interaction including hydrogen bonds, salt bridges, hydrophobic and van der Waals interactions. The program will define a box to cover interfacial residues of ligand protein and receptor protein, and create regularly spaced grids within this box. The grid spacing is 0.5 Å. It also defines residues of proteins in this box as pocket residues and atom in pocket residues as pocket atom. Three different types of probe atoms are used to screen those grids in this box which do not conflict with pocket atoms. The probes include (1) a positively charged sp^3 nitrogen atom (ammonium cation), representing a hydrogen bond donor, (2) a negatively charged sp^2 oxygen atom (as in a carboxyl group), representing a hydrogen bond acceptor and (3) an sp^3 carbon atom (methane), representing a hydrophobic group. Then for each grid, calculate three scores, hydrogen bond donor, hydrogen bond acceptor and hydrophobic, and judge the type of grid according to the three scores and those grids around it. Grid will be labeled as "donor", "acceptor", or "hydrophobic" according to the highest score on this grid and grids with no significant contribution will be filtered out. If atom in ligand lies in a "donor" grid, it could interact with a receptor as a hydrogen bond donor. So do "acceptor" grids and "hydrophobic" grids. These scores were calculated with the method developed by Wang et al [33]. Here is a brief description to the three scores. All the atoms on the protein pocket are labeled as either donor (D), acceptor (A), donor/acceptor (DA), or none (N). The bond length and bond angle are parameters for defining hydrogen bonds. In our program, we avoid the explicit use of hydrogen atoms in the structure. Therefore, we use the distance between grid and donor(D) or acceptor(A) atoms in protein pocket as a parameter to represent bond length. Use one angle involving only heavy atoms instead of the bond angle. It is among X–D(A)...grid, where X represents the adjacent heavy atom or, if there are more than one adjacent atom, their geometric centers. If the distance between grid and D(A) is shorter than the sum of vdW radii of standard acceptor(donor) atom and D(A) pocket atom, meanwhile, the angle X–D(A)...grid is more than 80, it is regarded as a hydrogen bond. Then count the number of hydrogen bond between this grid and pocket atoms. If the number is more than 4, select the four of them that have shortest distances. The hydrogen-bond score is the product of the number and a coefficient. The hydrophobic effect is related to the

Fig. 1 Key interaction grids deduced from the interface of barnase opposite to barstar. (The structure of barnase is represented by a ribbon. The hollow part in the surface of barnase is the binding site where key grids are deduced.) The figure was finished with Rasmol and texts were added in Photoshop 6.0

desolvation effect. We utilize a simple method to represent this effect. Each atom in a protein pocket is assigned a quantitative scale to represent it hydrophobicity [34]. The "environment" of a grid is as the assembly of all the neighboring protein atoms within 5 Å. The hydrophobicity of the environment is expressed by the sum of the hydrophobic scales of all the atoms forming the environment. If the sum is positive, the grid is considered to be in a hydrophobic environment and given a value (more than 0) as its hydrophobic score.

In the end, the characteristic of the interface is drawn out and output in PDB format. This characteristic exhibits key interaction sites of the acceptor's binding region. Users can look at it carefully and directly with graphic software such as Rasmol. As seen in Fig. 1, these sites line out of the van der Waals surface of receptor and the distances of these sites to receptor are consistent with the distances in which two atoms can non-covalently interact. In general, this distance is 2.0–3.6 Å for "donor" and "acceptor" grids and 2.5–5.0 Å for a "hydrophobic" grid. Considering that solvent may substitute for eliminated atoms of peripheral interfacial residues, so the program sets an exposing penalty (surface punishment) that atoms lying in the peripheral of the interface halve their contributions to the scores. We also need to mention that the user can select any part of protein as objects that are not necessarily close to the ligand. This means we can get all characteristics surrounding the protein. We extract key residues following a simple criterion—a key residue has enough atoms that can produce a cluster of key grids. Firstly, we count the number of coincident key grids surrounding an interfacial atom within a given distance. A large number means an important atom. Then the number of important atoms in each interfacial residue is counted. According to different scales of importance for the atoms and the number of important atoms, we divide interfacial residues into hot-spot residues, warm residues and unimportant residues, which can be compared directly to experimental data.

PP_SITE can be compiled under LINUX or UNIX. It is easy to use and the computational time depends on the size of the complex and interface concerned. The following example using barnase/barstar that have a buried interface about 1,300 $Å^2$, takes about 20 min on a SGI O2/R10000/150M. It takes only 4 min on a LINUX system with PIII600/256M if no optimization is used in compiling.

PP_SITE is now available freely via anonymous FTP from our ftp-server: ftp://mdl.ipc.pku.edu.cn/pub/ software/pp_site (or use its IP:162.105.177.40).

Results and discussion

Application of PP_SITE in protein–protein interface analysis

Firstly, we will give an example to explain the result the user can obtain. The interaction of barnase, an extracellular RNase of Bacillus anylolique-faciens, with its intracellular inhibitor barstar is a good example for protein–protein interaction study, as the structures of both the free and the complexed proteins are available at high resolution. And the existing mutation data is another advantage. We used barstar C(40,82)A mutant/barnase (PDB code 1B27 with resolution of 2.1 Å) [35] as the target to study. Figure 2 shows the characteristics of the interface.

Figure 2 gives a graphic display of the result at the interface. The places where hydrophobicity are strong and where hydrogen bonds may form can be seen clearly in the figure. According to [36, 37], Lys27, Arg59, Arg83, Arg87 and His102 of barnase are key residues for the

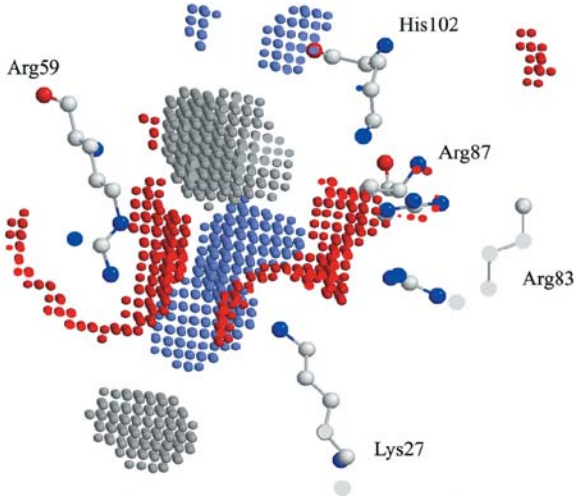

Fig. 2 The key interaction sites and key residues of barnase at the interface opposite to barstar. *White dots* represent hydrophobic sites; *red dots* represent hydrogen bond acceptor sites; sky blue dots represent hydrogen-bond donor sites. *White balls* represent carbon atoms; *blue balls* represent nitrogen atoms and *red balls* are oxygen atoms. Those labeled residues are the largest contributor to the binding according to [23]. The figure was drawn with Rasmol in depth mode in conjunction with slab mode. In Rasmol, the command depth or slab only draws those portions of the molecule that are closer to the viewer than a given *z*-clipping plane or further from the viewer than a given *z*-clipping plane. Integer values range from zero at the very back of the molecule to 100 which is completely in front of the molecule. Here we set value 40 for depth and value 50 for slab. Figures 3 and 5 are also treated like this. This operation can ensure that the figure is not a projection of all the atoms

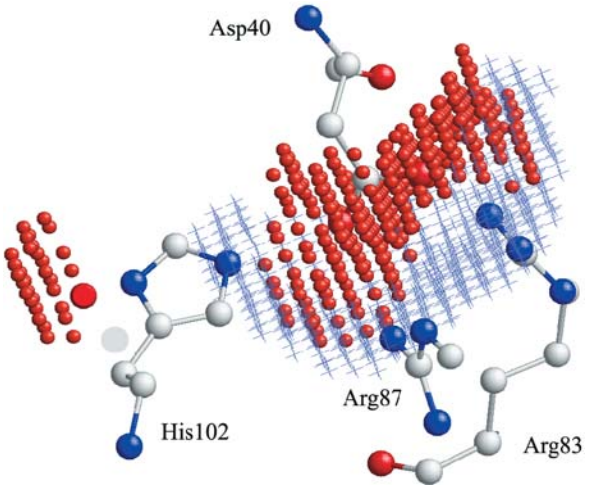

Fig. 3 Key interaction grids at the interface between barnase and barstar. Key interaction grids for barnase and barstar are presented in one figure according to their coordinates. In order to distinguish the two proteins, different icons were assigned. *Crosses* represent key grids for barstar and *dots* represent key grids for barnase. The four residues in ball and stick are Asp40 in barstar and Arg83, Arg87, His102 in barnase

Table 1 Comparison of predicted and experimental results of Ala mutation

Predicted	Experimental		
	H[a]	W[b]	N[c]
H	34[d] (38[e])	17 (19)	17 (25)
W	32 (26)	26 (27)	34 (31)
N	9 (11)	28 (25)	53 (48)

[a] H represents hot-spot residue
[b] W represents warm residue
[c] N represents unimportant residue
[d] Result with surface punishment
[e] Result with no surface punishment (in parentheses). For example, the experimental value H to the predicted value W denotes the number of hot-spot residues predicted to be warm residues

binding. In Fig. 2, they are those residues that make up an aggregation of key interaction sites. When the temperature factors in PDB files are replaced by the scores, commonly used graphic software can be used to analyze the magnitude of scores. On the other hand, we set up a parameter SURFACE_PUNISH in the program, and users can select if it is necessary to lower the influence of surface residues. If this parameter were not applied, Glu60 of barnase would also correspond to clusters of key interaction sites and be regarded as a hot-spot residue, but single mutation experiments showed that it does not influence the association energy too much. This may come from the fact that Glu60 lies on the surface of the protein and most of it interacts with solvent water, not with barstar. Trp44 and Trp38 of barstar are in the same situation as Glu60 of barnase.

In order to analyze the interactions, figures of the two key interaction sites for barnase and barstar were merged into one (Fig. 3).

As is well known, the protein–protein interface is complementary. Generally speaking, the complementarity should embody in hydrophobic sites produced by barnase near hydrophobic sites produced by barstar, hydrogen bond donor sites of barnase near hydrogen bond acceptor sites of barstar and vice versa. But in fact, distinct coincidence occurs at only one position: the central portion of the binding basin produced by Arg83, Arg87, His102/barnase and Asp40/barstar. In Fig. 3, we show only key sites deduced from these four residues just for bringing to a focus. From the data provided in [36], these residues had a significant cooperative effect in the double mutant cycles. Thus, this example shows that this kind of complementarity may be used as an indication of correlated mutations.

Two conclusions can be drawn from the above: (1) mutating those residues (that are not exposed to solvent) that make up a large aggregation of key interaction sites will change the association energy markedly (see Fig. 2). (2) If some of the key sites coincide with each other (see Fig. 3), those residues contributing to the sites may be cooperative. The change of association energies in a double-mutation cycle is not additive in these cases.

Comparison with alanine scanning mutation

We applied PP_SITE to 13 protein–protein complexes described in Materials and methods. The program exported key residues according to key grid distribution. Residues constituting a binding interface are divided into three classes: hot-spot residue, warm residue and unimportant residue. As mentioned before, hot-spot residues are those with $\Delta\Delta G \geq 1.5$ kcal mol^{-1}; warm residues, $1.5 > \Delta\Delta G \geq 0.5$ kcal mol^{-1}; unimportant residues, $\Delta\Delta G < 0.5$ kcal mol^{-1}. Table 1 shows the result of comparison between computational and experimental results. Alanine-scanning mutation reflects the influence of side chains on the stability of protein complexes. Thus side chain atoms on the periphery of interface are more easily replaced by water in a non-disruptive manner than atoms in the center of binding interfaces. So we introduced a punishment to the pocket residues on the surface of complex. The value in Table 1 is the result after running surface punishment and the value in bracket in Table 1 is the result before running surface punishment. In order to compare carefully, the detailed statistics of results were listed. In these 250 mutations, 75 are hot-spot residues, 71 are warm residues and 104 are unimportant residues. The results with surface punishment are a little better than with no punishment, as can be seen from Table 1. As only 9 out of 75 hot-spot residues were not predicted, the success rate for hot-spot residue prediction is 88%. When observed with graphic software, most of the hot-spot residues that were not predicted correctly

could not produce key grids, which might come from the conformational change introduced by alanine mutation. These conformational changes make real hot-spot atoms shift and cause large $\Delta\Delta G$. This may be the reason why the program could not find these hot-spot residues. It is also the blind corner for all the current methods as they cannot consider conformational change in mutation. We rank hot and warm residues as first class, unimportant residues as second class and the averaged successful rate of prediction is 65%. For the first class, the success rate is 75% and for the second class, the rate is 52%. From these results, we can see that this method overestimates unimportant residues. These falsely predicted unimportant residues partly lie at the periphery of interface, so this kind of error can be corrected through appropriate surface punishment. Corresponding to Table 1, this means that the smaller N/H and N/W, the better of the result. Comparing 1 and 2 of Table 1, there are eight unimportant residues predicted to be key residues with no surface punishment that are correctly predicted with surface punishment. The surface punishment has been shown to improve the results.

Tolerance to conformational changes

We compared the difference between key residues deduced from complexed and unbound structures. In the set of 26 protein–protein complexes, the total number of key residues deduced from complexed and unbound proteins is 449. In these key residues, there are 51 cases where the difference comes from absence of atoms in the crystal structure and 65 cases where the small difference comes from the simple judgement as the program imposes a uniform criterion on all situations when judging which residue is a key residue. The other two reasons are the conformational change of side chains and backbones, which account for 26 and 18 cases, respectively. Thus, the difference between results deduced from complexed and unbound structures is 11% ((26+18)/(449−51)). From the above we can see that most of the difference comes from the incompleteness of atoms in the crystal structure. The deformity of residue not only has an influence itself, but also affects other residues near it. This reminds us that we must check the crystal structure carefully when applying this program, especially for residues participating in binding.

With this test, we can say that PP_SITE can tolerant conformational change to a certain extent. It can be used in analyzing the unbound structures when the complexed structure is unknown.

Comparison with other methods

Massova and Kollman [20] have used molecular dynamics simulations to study the complex formation between tumor suppressor p53 and oncoprotein MDM2. The crystal structures used were 1ycr, 1ycq, which are human MDM2 complexed with residues 15–29 of human p53 and xenopus laevis MDM2 complexed with residues 13–29 of human p53. In experiments with phage display libraries, [38] the critical role of Phe19, Trp23 and Leu26 was emphasized, so that these three residues could not be replaced. Massova and Kollmann reproduced the qualitative trends in the experimental data for all 12 amino acids of the p53 peptide and identified four hydrophobic residues Phe19, Leu22, Trp23 and Leu26 as critical binding points. We have used PP_SITE for the same example. For 1ycr, we found that Glu17, Phe19, Leu22, Trp23 and Leu26 are hot-spot residues and Leu25, Pro27 and Asn29 are warm residues. For 1ycq, we found that Phe19, Leu22, Trp23 and Leu26 are hot-spot residues and Glu17 and Pro27 are warm residues. These results are consistent with experiment.

Verkhivker and Bouzida et al. (2002) [22]have used energy landscape analysis to study a 13-residue cyclic peptide DCAWHLGELVWCT binding to the Fc fragment of the Ig protein (PDB code 1dn2). They found that the most dramatic loss of binding affinity occurs when the Asn434, His433, His435 and Tyr436 residues were replaced by alanine. These results agree with mutagenesis experiments. [39] From our calculations with PP_SITE, Ile253 and Tyr436 are hot-spot residues and Met252, Glu382, Met428, Asn434 and His435 are warm residues in the Fc fragment of Ig protein, while His5, Leu6, Val10 and Trp11 are hot-spot residues and Asp1, Glu8 and Leu9 are warm residues in the peptide. This result is also consistent with experiment. [39]

Hu et al.'s (2000) [23] statistical analysis indicated that the percentages of the conserved residues range from 20 to 50% of all contacting residues with few exceptions. Except for a few outliers, the correlation coefficient of the experimentally determined amino acid enrichment and their computed conservation propensity is 0.72. In particular, conserved interface residues are strongly correlated with the experimentally identified hot spots, compiled from the database of experimental alanine-scanning mutagenesis. As Hu et al. (2000) [23] only listed the results of the 1choEI (serine proteinase inhibitor) and 1vfaAB (immunoglobulin) families, we compared our results from PP_SITE for these two families. The results are listed in Tables 2 and 3. For the 1cho family, Hu et al. predicted the conserved residues within four segments, 41–42, 57–58, 191–194 and 214–216. In our calculation, each of the four segments has a residue being predicted. For 1vfa family, Hu et al. predicted 36, 44, 46, 49, 95, 98 in the light chain and 39,45, 47, 94, 106, 107, 108 in the heavy chain to be conserved residues. Using one crystal structure, PP_SITE can correctly predict 11 out of the 13 conserved residues. Only residues 107 and 108 in the heavy chain cannot be predicted. Since conserved residues at the interface may contribute to structural stability, etc, not all of them are key residues for binding. This may be the reason why some of the conserved residues cannot be predicted by PP_SITE. In contrast, PP_SITE predicted close-space neighbors, such as 39, 99, 151, 192 in the 1cho family to be key sites that are not

Table 2 Result of PP_SITE applied to proteins in 1cho family

Res ID[a]	1cho[b] EI	1ppe[b] EI	1brc[b] EI	1ppf[b] EI	1tab[b] EI	1tgsZ[b] I	3sgb[b] EI	1mct[b] AI	2kai[b] BI	4tpiZ[b] I
35				Leu						
39	Phe*	Tyr*	Tyr*		Tyr*	Tyr*				Tyr*
41[c]	**Phe**			**Phe**			**Arg**	**Phe**		
57	**His**	**His**	**His**	**His**	**His**	**His**	**His**	**His**		**His**
64	Asp									
94	–						Phe			
99	Ile	Leu*	Leu*	Leu*	Leu*	Leu*		Leu*	Tyr*	Leu*
143	Leu*			Leu*						
146	Tyr*									
149		Thr	Val*						Phe*	
151	Thr	Tyr*		Ile*	Tyr*	Tyr		Tyr*	Phe*	Tyr*
152						Pro				
169							Val			
171							Tyr*			
172	Trp									
189		Asp			Asp	Asp		Asp	Asp	Asp
190								Ser		
192	Met*	Gln	Gln	Phe*		Gln	Pro*	Gln	Met*	Gln
195	**Ser**	**Ser**	**Ser**	**Ser**	**Ser**	**Ser**	**Ser**	**Ser**	**Ser**	**Ser**
215	**Trp**	**Trp***	**Trp***	**Phe***	**Trp***	**Trp***		**Trp***	**Trp**	**Trp***
217		Ser	Tyr*	Arg	Ser			Tyr*		
218	Ser									
226			Asp							
227							Phe			

[a] Res ID follows the residue id of 1cho
[b] * indicates hot-spot residues, others are warm residues predicted by PP_SITE
[c] Rows in bold are conserved residues according to Reference [12]. In Reference [12], the conserved residues are **41**, 42, **57**, 58, 191, 193, 194, **195**, 214, **215**, 216 for 1choEI family

conserved in the sequence. Residues at these positions may be the source of specificity of protein–protein interactions in the same family.

Other applications

Qualitatively compare the strength of interaction

The example is human immunodeficiency virus type 2 (HIV-2) protease complexed with its inhibitor Phe–Val–Phe–Psi (CH_2NH)–Leu–Glu–Ile–amide (BI-LA-398), whose PDB code is 2MIP [40] and resolution is 2.2 Å. We can see in the crystal structure that the ligand has two orientations, with an occupancy of 0.55 for chain E and 0.45 for chain G, shown in Fig. 4c. As the protease dimer has C_2 symmetry, the key interaction sites deduced from the interface of HIV-2 also have similar symmetry. The two orientations of the ligand fit well with key sites, so it is reasonable that this ligand can have two orientations. As can be seen from Fig. 4a and b that conformation one (chain E) fits better with the grids than conformation two (chain G). We may take this as a qualitative criterion to judge the strength of the binding by checking the percentage of atoms in the interface of ligand falling into the sites with the same type deduced from its partner.

Differentiate binding modes

The example is anti-hen egg white lysozyme antibody D1.3 complexed with hen egg white lysozyme(HEL) with PDB code 1VFB and D1.3 antibody complexed with the anti-lysozyme antibody E5.2 with PDB code 1DVF [37, 41, 42, 43]. We find a large difference between the key sites produced by HEL and E5.2. The key sites deduced from HEL are more disperse than those deduced from E5.2; the distribution of three kinds of key sites are different (see Fig. 5b). Thus, it is obvious that D1.3 binds HEL and E5.2 in very different ways. According to [42], D1.3 contacts with these two proteins essentially through the same set of combining site residues (and mostly the same atoms, so key sites deduced from that two D1.3 are very similar, as seen in Fig. 5a). However, a small subset of contact residues dominates the D1.3–HEL interaction, while D1.3–E5.2 interaction is dominated by a much larger subset [42]. This is consistent with our result. This example shows that our method can also be used to differentiate binding modes.

Protein–protein interfaces are rather complicated. Both the size of the buried interface area and the type of interaction are important factors influencing the complexity. Seen from PP_SITE, some interfaces have dispersive, smaller key sites just like HEL to D1.3 and some have centralized, bigger key sites just like barnase. This phenomenon may be related to the type

Table 3 Result of PP_SITE applied to proteins in 1vfa family

Res[a] ID	1vfa[b] AB	1fvc[b] AB	1rej[b] AB	1aag[b] LH	1fgv[b] LH	1fvb[b] LH	1hfm[b] LH	1igm[b] LH	1jhl[b] LH	1mig[b] LH	2fvw[b] LH
Light chain in FV fragment of antibody											
1											asp
30											Ile*
32	Tyr*				Tyr*	Tyr					Tyr*
34					Asn	His	His				
36[c]	**Tyr***	**Tyr***	**Tyr***	**Tyr***	**Tyr***	**Tyr***	**Tyr***	**Tyr***	**Tyr***	**Tyr***	**Tyr***
38	Gln	Gln		Gln	Gln	Gln		Gln		Gln	
42							Glu				
43	Ser			Pro		Ser					Ser
44	**Pro***	**Pro***	**Pro***	**Pro**	**Pro***	**Pro***	**Pro***	**Pro***	**Asn**	**Pro***	**Pro***
46	**Leu***	**Leu***	**Leu***	**Leu***	**Leu***	**Arg**	**Leu***	**Leu***	**Leu***	**Leu**	**Leu***
49	**Tyr***	**Tyr***	**Tyr***	**Tyr***	**Tyr***	**Tyr***	**Lys***	**Tyr***	**Tyr***	**Tyr***	**Tyr***
50	Tyr*			Trp*	Tyr*	Asp				Phe*	
55		Tyr*		His			Ile*	Glu		Glu	Phe
87	Tyr*	Tyr*	Tyr*	Tyr*	Tyr*	Tyr*	Phe*	Tyr*	Ile	Phe*	Tyr*
89	Gln	Gln	Gln			Gln	Gln		Gln		Leu
91	Phe*	His	Tyr	Tyr*		Trp*		Tyr*	His	His	
94		Thr	Leu*	Tyr*	Leu*	Asn	Trp*	Leu*	Tyr*		Gln
95			**Pro***	**Pro**	**Pro**			**Pro**	**Pro**		
96	Arg	Pro	Tyr*	Leu*	Pro*	Tyr*	Tye*	Leu*	Trp*	Arg	Leu*
98	**Phe***	**Phe***	**Phe***	**Phe***	**Phe***	**Phe***	**Phe***	**Phe***	**Phe***	**Phe***	**Phe***
Heavy chain in FV fragment of antibody											
33							Tyr*				
35					His						
37			Tyr*	Val			Ile			Val	
39	**Gln**	**Gln**	**Gln**	**Gln**	**Gln**	**Gln**	**Lys**	**Gln**			**Gln**
44				Arg							
45	**Leu***	**Leu***	**Pro***	**Leu***	**Leu***	**Leu***	**Leu***	**Leu***	**Leu***	**Leu***	**Leu***
47	**Trp***	**Trp***	**Leu***	**Trp***	**Trp***	**Trp***	**Tyr47***	**Trp***	**Trp***	**Trp***	**Trp***
50				Thr		Glu*	Tyr*		Asn		Glu*
52	Trp		Tyr*								
58			Gln	Phe*			Tyr*	Asp		Glu	
60									Asn		
61										Pro	Pro
94	**Tyr**	**Tyr***	**Tyr***	**Tyr***	**Tyr***	**Tyr**	**Tyr***	**Tyr**	**Tyr**	**Tyr**	**Tyr**
96			Gln								
98	Glu	Trp*	Tyr		Trp*	His		Arg	Asp	Tyr*	Leu
99									Asp		His
100		Phe*		Asp		Ser					Tyr*
101	Tyr*	Tyr*		Tyr*		Ser			Tyr*		Tyr*
102	Arg				Arg Tyr*[a]		Asp*	Val		Arg*	
103	Leu*	Met*	Leu*	Met*	phe*	Phe*		Leu* Phe*	Met*	Phe*	Met*
104		Asp	Pro*					Asp	Asp		
105		Tyr	Tyr*			Tyr*					
106	**Trp***	**Trp***	**Phe***	**Trp***	**Trp***	**Trp***	**Trp***	**Trp***	**Trp***	**Trp***	**Trp***

[a] Res ID follows the residue id of 1vfa

[b] * indicates hot-spot residues, others are warm residues predicted by PP_SITE

[c] Rows in bold are conserved residues according to Reference [12]. In Reference [12], the conserved residues are **36**, **44**, **46**, **49**, **95**, **98** in light chain and **39**,**45**, **47**, **94**, **106**, 107, 108 in heavy chain

of protein and is worth examining further in future studies.

Conclusion

From the above we can see that our structure-based method for finding key residues in protein–protein interactions has been proven to be successful in qualitatively identifying hot spots. It can find 66 from 75 hot-spot residues identified by alanine mutation in 13 protein–protein complexes. Some of the hot-spot residues not predicted correctly may not participate directly in binding, but their mutations bring shifts of interfacial structure to induce large change in binding free energy. Besides this kind of primary application, this program can also be

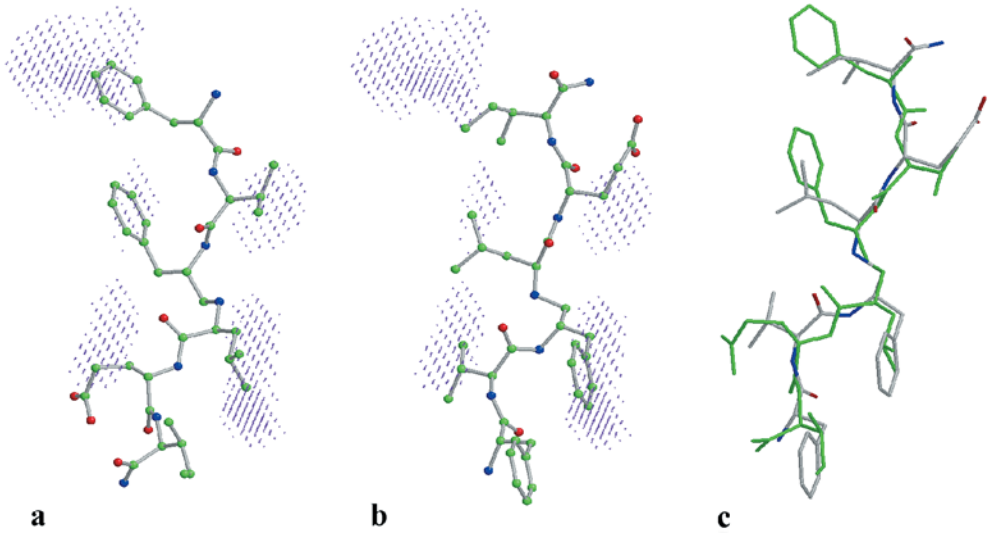

Fig. 4a–c Two orientations of HIV-2 protease inhibitor in crystal structure. These two orientations lay in the key grids produced by the HIV-2 protease binding site. Since the groups whose different orientations fall in hydrogen bond receptor and donor grids are similar, the figure only shows hydrophobic grids for clarity. **a** Orientation whose occupancy is 0.55. **b** Orientation whose occupancy is 0.45. **c** Two orientations in crystal structure (high occupancy orientation uses *green* color, low occupancy orientation uses *CPK* color)

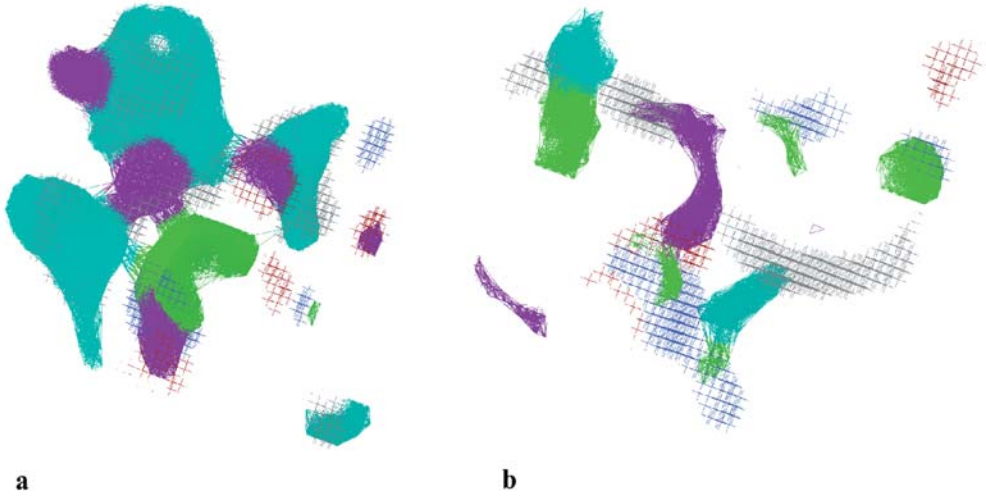

Fig. 5 a The key grids produced by D1.3 at the interface of D1.3/E5.2(1DVF) and D1.3/HEL(1VFB). **b** The key grids produced by E5.2 at the interface of D1.3/E5.2(1DVF) and by HEL at the interface of D1.3/HEL(1VFB). These two complexes were superimposed first. Key grids deduced from 1DVF are represented with crosses and colored with *CPK* mode. Key grids deduced from 1DVF are represented with patches and *cyan* for hydrophobic grids, *green* for donor grids, *magenta* for acceptor grids. **a** shows that two D1.3s result in very similar key grids, while **b** shows that E1.5 and HEL result in very different key grids

used to describe the interface with hydrogen-bond and hydrophobic properties, which makes the interactions on the interface easily observable. Furthermore, this method can analyze the interactions at the atomic level, which can help users know why a hot-spot residue is a hot spot. Thus, it can be used not only in alanine mutation analysis, but also in other mutational studies. The result can also be used as a direct beginning of drug design based on protein–protein interface.

Acknowledgement This work has been supported by the Ministry of Science and Technology of China (the 863 High-tech project and the Basic Research Project 2003CB715900), the National Natural Science Foundation of China and The Committee of Science and Technology of Beijing.

Appendix

Complexed				Unbound protein 1			Unbound protein 2		
PDB code	Res (Å)	Protein 1 chain	Protein 2 chain	PDB code	Res (Å)	Chain	PDB code	Res (Å)	Chain
Enzyme–inhibitor complex									
1brb	2.1	E	I	1bra	2.2	–			
1cgi	2.3	E	I	1chg	2.5	–	1hpt	2.3	–
2kai	2.5	A, B	I	2pka	2.1	A, B			–
2ptc	1.9	E	I	1bty	1.5	–	1bpi	1.1	–
2sic	1.8	E	I	1sup	1.6	–	3ssi	2.3	–
2sni	2.1	E	I				2ci2	2.0	–
1acb	2.0	E	I	5cha	1.7	A			
1brc	2.5	E	I				1aap	1.5	A
1cse	1.2	E	I	1scd	2.3	–			
1ppe	2.0	E	I				1lu0	1.03	A
1stf	2.4	E	I	1ppn	1.6	–			
1tgs	1.8	Z	I	1tgt	1.5	–			
2tec	2.0	E	I	1thm	1.4	–			
4htc	2.3	L, H	I	2hnt	2.5	–			
1udi	2.7	E	I	1udh	1.8	–	1ugi	1.55	A
Antibody–antigen complexes									
1mlc	2.1	A, B	E	1mlb	2.1	–	1lza	1.6	–
1vfb	1.8	A, B	C	1vfa	1.8	A, B	1lza	1.6	–
1nca	2.5	L, H	N				7nn9	2.0	–
1igc	2.6	L,H	A				1igd	1.1	–
2jel	2.8	L, H	P				1poh	2.0	–
Other complexes									
1atn	2.8	D	A	3dni	2.0	–			
1gla	2.6	G	F	1bu6	2.37	Y	1f3g	2.1	–
1spb	2.0	S	P	1sup	1.6	–			
2btf	2.6	P	A	1pne	2.0	–			
3hhr	2.8	A	B, C	1hgu	2.5	–			
1mda	2.5	L, H	A				1aan	2.0	–

Reference

1. Lichtarge O, Sowa ME (2002) Curr Opin Struct Biol 12:21–27
2. Bock JR, Gough DA (2001) Bioinformatics 17:455–460
3. Kini RM, Evans HJ (1995) Biochem Biophys Res Commun 212:1115–1124
4. Casari G, Sander C, Valencia A (1995) Nat Struct Biol 2:171–178
5. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) J Mol Biol 271:511–523
6. Gallet X, Charloteaux B, Thomas A, Brasseur R (2000) J Mol Biol 302:917–926
7. Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME, Lichtarge O (2002) J Mol Biol 316:139–154
8. Aloy P, Russell RB (2002) Proc Natl Acad Sci USA 99:5896–5901
9. Aloy P, Russell RB (2003) Bioinformatics 19:161–162
10. Zhou HX, Shan Y (2001) Proteins 44:336–343
11. Fariselli P, Pazos F, Valencia A, Casadio R (2002) Eur J Biochem 269:1356–1361
12. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N (2002) Bioinformatics 18 Suppl 1:S71–7
13. Wells JA (1991) Methods Enzymology 202:390–411
14. Clackson T, Wells JA (1995) Science 267:383–386
15. Bogan AA, Thorn KS (1998) J Mol Biol 280:1–9
16. Thorn KS, Bogan AA (2001) Bioinformatics 17:284–285
17. Jones S, Thornton JM (1996) Proc Natl Acad Sci USA 93:13–20
18. Lo Conte L, Chothia C, Janin J (1999) J Mol Biol 285:2177–2198
19. Elcock AH, Sept D, McCammon JA (2001) J Phys Chem B 105:1504–1518
20. Massova I, Kollman PA (1999) J Am Chem Soc 121:8133–8143
21. Huo S, Massova I, Kollman PA (2002) J Comput Chem 23:15–27
22. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Freer ST, Rose PW (2002) Proteins 48:539–557
23. Hu ZJ, Ma BY, Wolfson H, Nussinov R (2000) Proteins 39:331–342
24. Young L, Jernigan RL, Covell DG. (1994) Protein Sci 3:717–729
25. Villoutreix BO, Hardig Y, Wallqvist A, Covell DG, Frutos PG (1998) Proteins 31:391–405
26. Villoutreix BO, Covell DG, Blom AM, Wallqvist A, Friedrich U (2001) J Comput-Aided Mol Des 15:13–27
27. Goodford PJ (1985) J Med Chem 28:849–857
28. Gao Y, Wang RX, Lai LH (2002) Acta Phys-Chim Sin 18:676–679
29. Myers EW, Miller W (1989) Bull Math Biol 51:5–37
30. Delano WL (2002) Curr Opin Struct Biol 12:14–20
31. Betts MJ, Sternberg MJ (1999) Protein Eng 12:271–283
32. Wang RX, Gao Y, Lai LH (2000) J Mol Model 6:498-516
33. Wang RX, Liu L, Lai LH, Tang YQ (1998) J Mol Model 4:379–394
34. Wang RX, Gao Y, Lai LH (2000) Perspect Drug Discovery 19:47–66
35. Buckle AM, Chreiber GS, Fersht AR (1994) Biochem 33:8878–8889
36. Schreiber G, Fersht AR (1995) J Mol Biol 248:478–486
37. Covell DG, Wallqvist A (1997) J Mol Biol 269:281–297

38. Böttger A, Böttger V, Garcia-Echeverria C, Chène P, Hochkeppel HK, Sampson W, Ang K, Howard SF, Picksley SM, Lane DP (1997) J Mol Biol 269:744–756
39. DeLano WL, Ultsch MH, de Vos AM, Wells JM (2000) Science 287:1279–1283
40. Tong L, Pav S, Pargellis C, Do F, Lamarre D, Anderson PC (1993) Proc Natl Acad Sci USA 90:8387–8391
41. Goldman ER, Dall'Acqua W, Braden BC, Mariuzza RA (1997) Biochem 36:49–56
42. Dall'Acqua W, Goldman ER, Eisenstein E, Mariuzza RA (1996) Biochem 35:9667–9676
43. Dall'Acqua W, Goldman ER, Lin W, Teng C, Tsuchiya D, Li HM, Ysern X, Braden BC, Li YL, Smith-Gill SJ, Mariuzza RA (1998) Biochem 37:7981–7991