JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

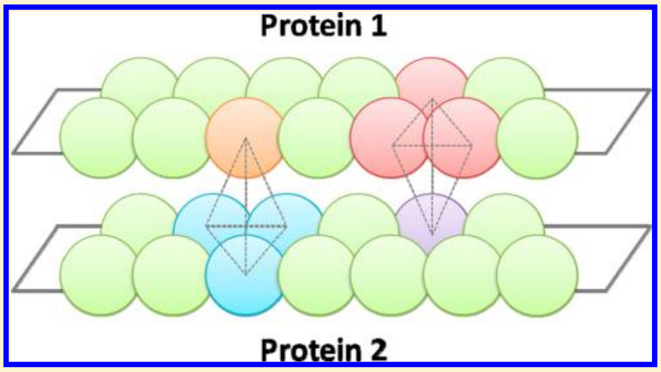
Mining the Characteristic Interaction Patterns on Protein−Protein Binding Interfaces

Yan Li,[†] Zhihai Liu,[†] Li Han,[†] Chengke Li,[†] and Renxiao Wang*,[†,‡,§]

[†] State Key Laboratory of Bioorganic and Natural Products Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, 345 Lingling Road, Shanghai 200032, People's Republic of China

[‡] State Key Laboratory of Quality Research in Chinese Medicine, Macau Institute for Applied Research in Medicine and Health, Macau University of Science and Technology, Macau, People's Republic of China

Supporting Information



ABSTRACT: Protein−protein interactions are observed in various biological processes. They are important for understanding the underlying molecular mechanisms and can be potential targets for developing small-molecule regulators of such processes. Previous studies suggest that certain residues on protein−protein binding interfaces are ″hot spots″. As an extension to this concept, we have developed a residue-based method to identify the characteristic interaction patterns (CIPs) on protein−protein binding interfaces, in which each pattern is a cluster of four contacting residues. Systematic analysis was conducted on a nonredundant set of 1,222 protein−protein binding interfaces selected out of the entire Protein Data Bank. Favored interaction patterns across different protein−protein binding interfaces were retrieved by considering both geometrical and chemical conservations. As demonstrated on two test tests, our method was able to predict hot spot residues on protein−protein binding interfaces with good recall scores and acceptable precision scores. By analyzing the function annotations and the evolutionary tree of the protein−protein complexes in our data set, we also observed that protein−protein interfaces sharing common characteristic interaction patterns are normally associated with identical or similar biological functions.

1. INTRODUCTION

Protein−protein interactions (PPI) play important roles in various biological processes, such as cellular regulation, signal transduction, DNA replication, protein synthesis and secretion, immune responses, and virus packing.[1−3] Execution of these complicated processes requires specific recognitions between different proteins. Elucidating the fundamental biophysical principles in protein−protein interactions is therefore of primary importance for the basic research in life science.

It is well-known that a protein−protein binding interface is quite different from a protein−ligand binding interface in several aspects. First, a protein−protein binding interface usually has a much larger contacting surface (1500−3000 Å^2)[4] as compared to a typical protein−ligand binding interface (300−1000 Å^2).[5] Second, a protein−protein binding interface is relatively flat, whereas a protein−ligand binding interface often has a well-defined cavity for hosting small-molecule binders. Yet, high affinity and specificity can be achieved in protein−protein binding. How is a specific recognition possible on a large and flat binding interface? Some sorts of distinctive interactions certainly participate in this process. Therefore, mining such interaction patterns on protein−protein binding interfaces will help to interpret the molecular mechanisms of protein−protein recognition and may have useful applications to protein engineering, drug design, and so on.

A number of studies have engaged on the statistical analysis of protein−protein interactions. In early years, researchers derived the preferences of residue−residue pairs on protein−protein binding interfaces, such as the pioneering work by Yan and Ofran groups.[6,7] In these studies, mathematical and statistical methods are applied to derive the preference of each possible residue pair. By comparing to the references of different residue pairs on binding interfaces, dominant residue pairs can be derived. This type of studies is of course a very straightforward interpretation of protein−protein complex structures. However, they do not provide much information beyond pairwise interactions. They can be misleading in some cases since recognition between protein molecules often occurs in a cooperative manner. This type of studies cannot deduce the possible relationship between protein−protein complex structures and their biological functions either.

The second type of studies aims at identifying ″hot spots″, which refer to the key residues which affects the binding process of a protein and its partner protein.[8,9] For example, the

Received: April 25, 2013
Published: August 10, 2013

algorithm developed by Tuncbag and co-workers[10] considered three features to predict hot spots, including conservation, accessible surface area, and statistical pairwise potentials of the residues on binding interface. The hot spot residues predicted by their algorithm matched the experimentally verified hot spots in the Binding Interface Database (BID)[11] with an average recall score of 70% and a precision score of 73%. A recently published work by Kozakov described a computational solvent mapping method for determining ″druggable″ hot spots on protein−protein binding interface.[12] They adopt a so-called FTMAP algorithm[13] to perform a global search over the entire protein surface for regions that favor the binding of certain ″probes″, i.e. some small organic molecules. Overlapping clusters of different probes are defined as the consensus sites. Then, the largest consensus site is predicted to be the most important hot spot in the protein binding site and so on. So far, a number of computational approaches for hot spot prediction have been developed. Several of them offer online computation service, such as Robetta,[14] HotPoint,[15] KFC,[16] PRICE,[17] and PCRPi-W.[18]

The third type of studies attempts to identify geometrically conserved patterns in protein−protein interactions. For example, Mintz and co-workers published their method for mining the similar interaction patterns in protein−protein interactions in 2005.[19] Their method can be described briefly as follows. The functional groups in each residue were simplified as ″pseudocenters″. Five types of 'pseudocenters' are defined, including hydrogen bond donor, hydrogen bond acceptor, mixed hydrogen bond donor and acceptor, hydrophobic aliphatic center, and hydrophobic aromatic center. With these definitions, a given protein−protein binding interface can be simplified into a set of pseudocenters. Geometrical alignment and comparison of different protein−protein binding interfaces were carried out using the I2I-SiteEngine algorithm.[20,21] The protein−protein binding interfaces with high matching scores were found to share similar interaction patterns. Protein−protein complexes were finally grouped into clusters by the matching scores of their binding interfaces. Functional relationships were observed among the protein−protein complexes grouped in the same cluster. Another derivative study was also published by the same group in 2007.[22]

We prefer the third type of studies, because they attempt to characterize protein−protein binding interfaces with three-dimensional descriptors. Our study followed this approach but developed a very different method. The basis of our method is to use residue clusters rather than individual residues or residue pairs to represent the key factors in protein−protein interaction. Technically, an interaction pattern by our definition is a cluster of four contacting residues: One residue, i.e. the ″anchor″, is on one protein molecule; while the other three are on another protein molecule (Figure 1). Different protein−protein binding interfaces can be compared by the composition of such interaction patterns. An intermolecular interaction is often more stable if they are formed within an appropriate environment, which could provide geometrical constraints, a low-dielectric shield from solvent, and so on. A cluster of residues has to be used to describe such an environment. It is thus better to rely on such three-dimensional interaction patterns for analyzing protein−protein interactions. Remember that a genetic code (codon) is composed in three nucleotides for some good reasons. If there is a hidden layer between the primary amino acid sequence and the tertiary structure of a protein molecule for coding its function, we hope that those codes are composed in clusters of reserved residues.

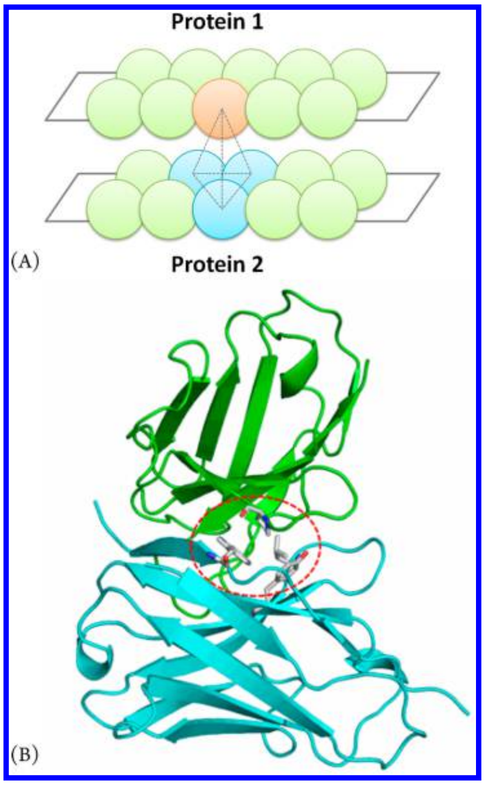


Figure 1. (A) Illustration of our residue-based definition of ″interaction pattern″ on a protein−protein binding interface. (B) A real example of interaction pattern (P-LYW) formed between two binding protein molecules (PDB entry: 1FVC).

In our study, a set of over 4,400 protein−protein binding interfaces retrieved from the Protein Data Bank (PDB)[23] was analyzed to sample the so-called interaction patterns. The interaction patterns conserved across different protein−protein binding interfaces were referred to as characteristic interaction patterns (CIPs). Similar to the concept of ″pharmacophore″ in drug design, we assume that such characteristic interaction patterns are the key factors for maintaining the biological function of a protein−protein complex. Our survey on the biological functions of the protein−protein complexes included in our data set supported this assumption. Based on the definition of characteristic interaction patterns, our method is also able to predict the ″hot spot″ residues on a protein−protein binding interface with good recall scores and acceptable precision scores on multiple data sets. Detailed descriptions and discussion of our method are given in the following sections.

2. METHODS

2.1. Preparation of Data Set.

A data set consisting of protein−protein complexes was compiled based on the entire Protein Data Bank (PDB). In order to distinguish valid protein−protein binding interfaces from those as the result of crystal packing, we adopted the method described by Tsai[20] with some necessary modifications. Four criteria were applied here: (1) A valid protein−protein binding interface is composed of two peptide chains from two individual protein molecules. (2) If the distance between a pair of heavy atoms, each of which locates on one protein molecule, is shorter than the sum of their van der Waals radii plus 0.5 Å, the two amino

Table 1. Definition of the Feature Centers on 20 Amino Acid Residues

category[a]	description	residue	location of the feature center
0	residues with a small side chain	Gly (G)	α-carbon atom on the backbone
		Ala (A)	β-carbon atom on the side chain
1	residues with a bulky hydrophobic side chain	Val (V)	β-carbon atom on the side chain
		Leu (L)	γ-carbon atom on the side chain
		Ile (I)	geometric center of β-, γ1-, γ2-, and δ-carbon atoms on the side chain
		Cys (C)	γ-sulfur atom on the side chain
		Met (M)	δ-sulfur atom on the side chain
2	residues with an aromatic side chain	Phe (F)	geometric center of the phenyl ring
		Tyr (Y)	geometric center of the phenol ring
		Trp (W)	geometric center of the indole ring
3	residues with a side chain that has a hydroxyl group	Ser (S)	γ-oxygen atom on the side chain
		Thr (T)	γ-oxygen atom on the side chain
4	residues with a side chain that has a carboxylic acid group	Asp (D)	γ-carbon atom on the side chain
		Glu (E)	δ-carbon atom on the side chain
5	residues with a side chain that has an amide group	Asn (N)	γ-carbon atom on the side chain
		Gln (Q)	δ-carbon atom on the side chain
6	residues with a side chain that has a positively charged group	Arg (R)	ζ-carbon atom on the side chain
		Lys (K)	ζ-nitrogen atom on the side chain
7	all histidine residues	His (H)	geometric center of the imidazole ring
8	all proline residues	Pro (P)	geometric center of the pyrrole ring

[a] Amino acid residues belong to the same category are considered as degenerate members.

acid residues containing these two atoms are considered as interfacial residues. The assembly of such residues forms the binding interface. (3) Any residue other than the 20 standard residues or structurally incomplete residues are not considered. (4) If a binding interface is composed of fewer than 10 residues, it is not considered as a valid protein−protein binding interface.

Three-dimensional structures of protein−protein complexes were all downloaded from PDB. By January 2012, a total of 78,235 experimentally resolved structures were deposited in PDB. Binary protein−protein complexes among them were identified using an in-house computer program. A total of 7,700 valid protein−protein binding interfaces were retrieved using the criteria described above. An additional criterion applied here was that the structures with overall resolution worse than 2.5 Å were not considered. The structures resolved with NMR or EM techniques are not considered in our data set, either. Next, we also discarded the protein−protein complexes composed with two identical chains. After these preparations, a total of 4,490 protein−protein binding interfaces were finally selected. Each binding interface was annotated after the original PDB code and peptide chain labels. For example, binding interface 1A2X-AB is denoted for the binding interface formed by peptide chain A and peptide chain B in the protein−protein complex structure in PDB entry 1A2X. The composition of each binding interface and structural coordinates were saved in a uniform format for subsequent analyses.

These protein−protein complexes were further clustered to remove redundant samples. For this purpose, these complexes were clustered by sequence similarity, which was computed with the BLAST program (version 2.2.20).[24] Because two peptide chains are involved in one binding interface, each protein−protein complex has two sequence similarity scores when being compared to a given peptide chain. In our study, the higher similarity score was considered in the subsequent clustering process. Three different levels of sequence identity cutoffs (50%, 70%, and 90%) were considered in clustering. With these cutoffs in clustering, the 4,490 protein−protein binding interfaces included in our data set were clustered into 895, 1016, and 1222 clusters, respectively. Then, the complex with the maximal number of contacting residue pairs was selected out as the representative of each cluster. Thus, our final data set includes three sets of protein−protein complex clusters, which will be referred to as “set-50”, “set-70”, and “set-90” throughout this article. Since one complex was selected out of each cluster, these three data sets consist of 895, 1016, and 1222 selected protein−protein complex structures, respectively.

2.2. Definition of Interaction Patterns.

Amino acid residues are the basic building blocks of protein molecules. It is thus convenient to describe the interaction patterns on a protein−protein binding interface with a residue-based algorithm. A residue may interact with multiple residues on another peptide chain. In our method, only the three nearest contacting residues were considered. Note that these three residues do not need to be continuous on the primary sequence. The anchor residue and other three residues form a tetrahedron (Figure 1). The reason for considering residue clusters is because we believe that critical interactions between protein molecules exist as three-dimensional patterns rather than one-dimensional residue pairs, and a tetrahedron is the minimal solution for defining a three-dimensional shape.

In order to quantify the distance between two contacting residues, each residue was represented by a certain “feature” center (Table 1). The distance between two residues was then computed as the distance between the two corresponding feature centers. The feature center of each type of residue was assigned to an appropriate atom on its side chain, which is typically the most representative part of this residue.

2.3. Preference of Interaction Patterns.

For a given protein−protein binding interface, a set of different interaction patterns can be derived. Which interaction patterns play critical roles? Are some interaction patterns actually conserved across different protein−protein interfaces? To answer these questions, we analyzed the preferences of interaction patterns. A preference factor (PF) was introduced to this analysis, which is calculated with the following equation:

$$PF = \frac{N_{\text{pattern\_}i}/\sum_i N_{\text{pattern\_}i}}{(N_{\text{res1}}/\sum_j N_{\text{res}j})(N_{\text{res2}}/\sum_j N_{\text{res}j})(N_{\text{res3}}/\sum_j N_{\text{res}j})(N_{\text{res4}}/\sum_j N_{\text{res}j})} \quad (1)$$

In this equation, $N_{\text{pattern\_}i}$ is the occurrence of interaction pattern i in a data set, and $\sum_i N_{\text{pattern\_}i}$ is the total occurrence of all types of interaction patterns in this data set. N_{res1}, N_{res2}, N_{res3}, and N_{res4} are the occurrences of four residues which form the given interaction pattern i, respectively, and $\sum_j N_{resj}$ is the total occurrence of all residues in this data set. Note that all 20 natural amino acid residues were classified into nine degenerate categories by the nature of their side chain groups (Table 1).

A preference factor value was assigned to each type of interaction pattern based on statistical analysis. A higher preference factor indicates that this interaction pattern is preferred more on protein−protein binding interfaces. Since three different sequence identity cutoffs (i.e., 50%, 70%, and 90%) were used in protein−protein interface clustering, the final cluster numbers in the three data sets were different. Thus, some interaction pattern types were found in Set-90 but not Set-70 or Set-50, and the preference factors of such patterns were of course different on three data sets. As for the interaction patterns included in all three data sets, their preference factors are generally consistent. In addition, preference factors were also used to identify the interaction patterns shared by different protein−protein interface. Applications of preference factors will be described in the following sections.

2.4. HOT SPOT INDEX

We assume that ″hot spot″ residues tend to be included in conserved interactions patterns. Here, we define a hot spot index to measure the probability of a certain residue as hot spot residue:

$$\textit{Hot spot index} = f_{\text{res}i} \times \sum_j PF_{\text{pattern\_}j} \quad (2)$$

Here, $\sum_j PF_{\text{pattern\_}j}$ is the sum of the preference factors of all interaction patterns containing residue i. $f_{\text{res}i}$ denotes for the occurrence probability of residue i on all protein−protein binding interfaces included in the data set. The product of them is the overall contribution of residue i. A larger hot spot index indicates a higher probability for residue i to be a hot spot residue.

Two external data sets were used as the training set and the test set to validate this method. The training set contains 349 residue mutation data obtained on 20 protein−protein complexes. Among these residues, 81 of them are known to be hot spot residues (experimental $\Delta\Delta G \geq 2.0$ kcal/mol). The test set contains 41 residue mutation data obtained on nine protein−protein complexes, including ten hot spot residues. The mutation data of both data sets were cited from the work published by David and co-workers.[25] In deriving the hot spot index, the probability of each residue type was computed over all protein−protein binding interfaces in Set-90. The final results are summarized in Tables S3 and S4 in the Supporting Information.

Precision scores and recall scores are two popular indicators for evaluating the performance of predictions. Precision score is defined as TP (true positive) divided by the sum of TP and FP (false positive); whereas recall score is defined as TP divided by the sum of TP and FN (false negative). In our study, precision and recall scores were computed under several cutoffs of hot spot index. David's transductive support vector machine (TSVM) method[25] and the Robetta method[26] were also applied to the same data set for comparison.

Our method was also tested on the Bcl-2 family proteins as a demonstration of its ability for predicting hot spots. This protein family consists of key regulators of the intrinsic apoptotic pathway. Both the pro-apoptotic and antiapoptotic members of this family execute their antiapoptotic functions through protein−protein interactions.[27−30] There are abundant residue mutation data obtained for Bcl-2 family proteins as well as their peptide ligands. Here, we collected 23 single-mutation data for three protein−protein complexes, i.e, Bcl-x$_L$/Bak (PDB entry: 1BXL),[31] Bcl-x$_L$/Bad (PDB entry: 1G5J)[32] and Bcl-2/Bax (PDB entry: 2XA0).[33] Interaction patterns on the binding interface were retrieved with our method based on the known three-dimensional structures of these complexes. The hot spot index of each mutated residue was calculated by eq 2. Performance of our method was then evaluated by comparing the computed hot spot indices with experimental mutation data.

2.5. Comparison of Protein−Protein Complexes Sharing Common Interaction Patterns. For the interaction patterns shared across different protein−protein binding interfaces, we examined the intrinsic relationships between them to investigate whether they are conserved. To achieve this goal, another parameter called "conservation score" is defined in eq 3 to measure the level of conservation of an interaction pattern between a pair of protein−protein binding interfaces.

$$\textit{Conservation Score} = PF(pattern\_n) \times Sim_{ij}/RMSD_{ij} \quad (3)$$

Here, $PF(pattern\_n)$ is the original preference factor of the shared interaction pattern between two protein−protein binding interfaces. Considering that the residues in the same degenerate category still have minor differences (e.g., Val and Leu), if all four residues are not identical between a pair of interaction patterns, the difference might be enlarged unreasonably. To overcome this problem, the Tanimoto similarity (Sim_{ij}) is also introduced into this equation to reflect the chemical similarity between two interaction patterns. $RMSD_{ij}$ is used to reflect the geometrical difference between two interaction patterns, which is calculated as the root-mean-squared-deviation (in Angström) of the four member residues between two interaction patterns after they are superimposed. In our study, superimposition of two interaction patterns was performed with the "RMS_FIT" program implemented in the AMBER software suite (version 9).[34]

The conservation score was used to group different protein−protein binding interfaces into clusters under certain cutoffs. All of the members in each cluster were manually checked whether they have identical or similar biological functions or belong to the same protein family through the annotations available on the PDB Web site. If this was confirmed, then the clustering was considered to be correct. By monitoring the rates of correct clustering under different conservation score cutoffs, an optimal cutoff value for grouping different protein−protein complexes with identical or similar biological functions could be determined.

3. RESULTS AND DISCUSSION

3.1. Interaction Patterns and Their Preferences. The protein−protein binding interfaces included in Set-50, Set-70, and Set-90 were examined by a set of computer programs to retrieve interaction patterns. The interaction pattern types and corresponding occurrences were counted to compute the probability of each type of interaction pattern. Occurrences of the residues participating in forming the interaction patterns were also counted to calculate their background probabilities on all protein−protein binding interfaces. Then, preference factors of all interaction patterns were computed with eq 1 for all three data sets (see Table S2 in the Supporting Information). Due to the different sizes of three data sets, the final probabilities were not identical. For the sake of convenience, the results obtained on Set-90 are discussed below since this data set provides a wider coverage of different protein−protein complexes.

By our method, all 20 natural amino acid residues are grouped into nine categories by their side-chain properties (Table 1). An interaction pattern is coded after the categories of four member residues. For example, an interaction pattern formed by Phe, Gly, Val, and Thr is coded as “3-0-1-2”. Given nine categories of residues, there are a total of 1,485 types of interaction patterns in theory. Our results indicate that Set-90 contains 1,375 of them. Among the 110 missing types, most of them have two histidine (category #7) or proline residues (category #8) simultaneously, e. g. “0-1-7-7″, “4-5-7-7″, “0-6-8-8″, “5-6-8-8”, and so on. Although some patterns with two proline residues are actually observed, their preference factors are very low. For example, the preference factors of pattern “1-4-8-8” and “4-2-8-8” are 0.41 and 0.85, respectively. In other words, the probability of finding interaction patterns composed of two histidine or proline residues at protein−protein binding interfaces is lower than the background reference, i.e. a combination of four random residues.

Distribution of the preference factors of all interaction pattern types derived from Set-90 are shown in Figure 2. One can see that the distribution peak locates between 3.0 and 4.0. When the preference factor is larger than 7.0, the corresponding occurrence decreases dramatically below 100. Only a few interaction pattern types have a preference factor higher than 20, such as “3-0-3-8” (*PF* = 24.7), “3-2-3-8” (*PF* = 25.3), “5-7-7-7” (*PF* = 31.4), and so on. These high-PF patterns usually include residues in category “5″, “7”, and “8” (Table 1) since these three residue types have relatively low background probabilities. According to our definition, a larger preference factor indicates that an interaction pattern is more favored on protein−protein binding interfaces. It is reasonable to expect that such favored interaction patterns are crucial for the binding of two partners. For example, the interaction pattern showed in Figure 1B is Pro-Leu-Tyr-Trp, with a preference factor of 17.1. Indeed, this Trp residue was reported to be a hot spot on the binding interface formed by the heavy chain and the light chain of humanized anti-p185HER2 antibody 4D5.[35]

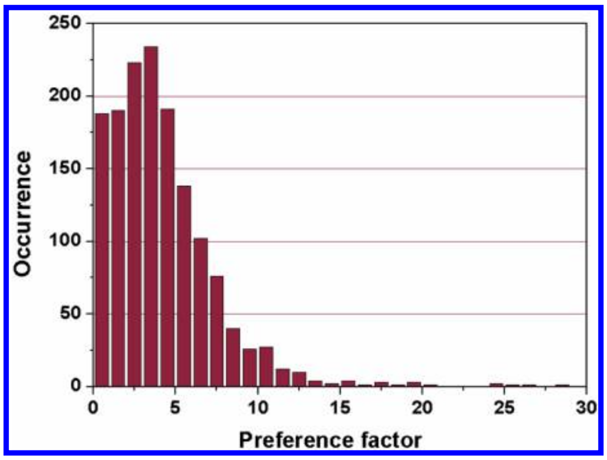


Figure 2. Distribution of the preference factors of the interaction patterns derived from all of the protein−protein binding interfaces in Set-90.

Our method is very different from other methods for analyzing protein−protein binding interfaces. For example, in Mintz’s method[19] each residue is represented by a pseudocenter, which has five categories in total. Then, two protein−protein binding interfaces can be compared by comparing these two sets of pseudocenters. In contrast, the basic element in our method is the so-called interaction pattern, each of which is a cluster of four residues, whereas other methods are based on either individual residues or residue pairs. In our method, the 20 natural amino acid residues are classified into nine categories by the properties of their side chains, making it convenient to compare two interaction patterns directly based on their compositions.

It also should be explained why an interaction pattern in our method consists of four residues: Three residues are the minimal set for defining a patch on the surface of a protein molecule, and four residues are the minimal set for defining a three-dimensional shape between two protein molecules. It is also possible to expand an interaction pattern to include more residues. In that way, the interaction pattern will be able to include even more structural details. However, this practice also has certain technical problems. First, the total number of possible interaction patterns will increase dramatically, and this will lead to a much lower occurrence of each individual interaction pattern. Consequently, it will become difficult to derive any significant conclusions through statistical analysis. Second, if an interaction pattern consists of more residues, it will become complicated to describe its geometrical shape or to make comparison of two patterns. Due to these reasons, we did not attempt to go beyond four residues for defining the interaction pattern.

3.2. Prediction of Hot Spots.

In our method, two interaction patterns are allowed to overlap. In other words, the same residue may be the member of two or more interaction patterns. If a residue is found in multiple favored interaction patterns (i.e., those with high preference factors), we assume that it is likely to be a hot spot residue. Then, an algorithm for predicting the hot spot residues on protein−protein binding interface can be designed based on this assumption. In our study, each protein−protein complex in the training set and the test set was analyzed to derive a set of all possible interaction patterns on the binding interface. Each residue in these two sets was assigned a hot spot index according to eq 2. If a residue was not found in any interaction pattern, its hot spot index was set to zero. Then, three different cutoffs, i.e. 0.30, 0.40, and 0.50, are used to classify these residues into hot-spot residues and nonhot-spot residues. The final statistical results are listed in Table 2, together with the results provided by David in their literature[25] and the predictions made by the Robetta method.[26]

In the case of our method, when lower cutoffs of the hot spot index were applied to the classification of the training set, it resulted in decreased precision scores and increased recall scores. The same trend was also observed in the results produced by the Robetta method. When the cutoff was 0.30, our method produced a good recall score of 0.72, which is very

Table 2. Performance of Three Methods in Hot Spot Prediction on the Training Set and the Test Set

	training set		test set	
method	precision score	recall score	precision score	recall score
hot spot index (0.50)[a]	0.45	0.56	0.43	0.60
hot spot index (0.40)	0.40	0.63	0.35	0.60
hot spot index (0.30)	0.36	0.72	0.33	0.70
Robetta (2.0)[b]	0.52	0.47	0.42	0.50
Robetta (1.8)	0.53	0.52	0.36	0.50
Robetta (1.0)	0.39	0.75	0.32	0.70
TSVM	0.56	0.65	0.31	0.40

[a]Three cutoffs, i.e. 0.30, 0.40, and 0.50, were used in our classification of hot-spot residues and nonhot-spot residues. [b]Three cutoffs, i.e. 1 kcal/mol, 1.8 kcal/mol, and 2.0 kcal/mol, were used in Robetta's classification. Results of the Robetta method were produced by the online Robetta server (http://www.robetta.org/).

Table 3. Experimental Mutation Data and Predicted Hot Spot Residues on Bcl-2 Family Proteins for Binding with BH3 Peptides

mutations		K_d (FP, μM)[a]	$\Delta\Delta G$ (kcal/mol)[b]	hot spot index[c]	Robetta (kcal/mol)[d]
Bcl-x$_L$/Bak (1BXL)	WT	0.34	—	—	—
	V574A	15	2.26	1.302	1.18
	R576A	3.3	1.35	0.260	0.27
	L578A	270	3.98	2.585	2.67
	I580A	1	0.64	0.218	0.51
	I581A	17	2.33	0.676	2.17
	G582A	0.5	0.23	0	—
	D583A	41	2.86	0	0
	D584A	0.14	−0.53	0	−0.03
	I585A	93	3.34	0.597	1.07
Bcl-x$_L$/Bad (1G5J)	WT	0.0006	—	—	—
	N301A	0.0004	−0.24	0	−0.03
	L302A	0.0007	0.92	0	1.81
	W303A	0.0003	−0.41	0	0.18
	A304G	0.0008	0.17	0	0.01
	A305G	0.0024	0.83	0.424	−0.01
	S322A	0.0003	−0.41	0	−0.03
	F323A	0.0021	0.75	0.510	2.96
	K324A	0.0012	0.41	0	0
	K325A	0.0002	−0.65	0	0
Bcl-2/Bax (2AX0)	WT	0.0151	—	—	—
	E61A	0.0952	1.10	0	0.28
	R64A	0.129	1.28	0.968	1.86
	D68A	1.04	2.52	0.500	3.42
	E69A	0.476	2.06	0.625	2.27
	R78A	0.0571	0.79	0.689	0.74

[a]Dissociation constants of the Bcl-x$_L$/Bak, Bcl-x$_L$/Bad, and Bcl-2/Bax complexes are cited from refs 31, 32, and 33, respectively. [b]ΔG values are computed from experimentally measured dissociation constants. The residues with $\Delta\Delta G > 2.0$ kcal/mol are generally considered as hot spots. [c]Computed by our method with eq 2. [d]Results of the Robetta method, which were given by the online Robetta server (http://www.robetta.org/submit.jsp).

close to the Robetta result under a cutoff of 1.0 kcal/mol, and also better than the TSVM result, 0.65. However, the precision score of our method is slightly poorer on the training set, indicating that our method predicted more false positives. Nevertheless, our method gives much better performances than the TSVM method on the test set, with the precision scores in the range of 0.33−0.43 and recall scores in the range of 0.60−0.70 under all three cutoffs. These results are also comparable or marginally better than the Robetta method (Table 2). The obvious poorer result of th TSVM method on the test set suggests it is highly dependent on the training set. This is the common limitation of most machine learning methods. Differently, our method has little effect on the performances for different data sets since it is not developed based on either of them. The preference factor and residue probability required in our method are both derived from protein−protein binding interfaces in Set-90. Therefore, the improvement of this external data set will finally increase the performance of our method. At the current stage, our method has some drawbacks in distinguishing between true and false positives. It is what we need to improve in next plans. Anyway, the high recall score indicates that we can still apply this prediction method in a small scope with much less cost.

We then tested our method on the complexes formed among Bcl-2 family proteins, which is a well-known protein−protein interaction system. The computed hot spot index of each mutated residue on these three complexes is summarized in Table 3. Also listed in this table are the $\Delta\Delta G$ energies of all mutated residues predicted by the online Robetta server.[26] Computed values were compared to the experimental mutation data to judge if a prediction is correct or not. Under three different cutoffs (0.30, 0.40, and 0.50), our method produced precision scores of (0.60, 0.60, 0.67) and recall scores of (0.86, 0.86, 0.86), respectively. The Robetta method produced precision scores of (0.67, 0.57, 0.80) and recall scores of (0.86, 0.57, 0.57) under three $\Delta\Delta G$ cutoffs (1.0 kcal/mol, 1.8 kcal/mol, and 2.0 kcal/mol). One can see that our method exhibited a more consistent performance under different cutoffs in this test case.

We then examined the false positives and false negatives made by our prediction on the three complex structures. When the cutoff for the hot spot index is set as 0.40, there are two false positives on the Bcl-2/Bax complex, i.e. Arg64 and Arg78. Although these two residues indeed form salt bridges with their partner residues (Figure 3A), their corresponding mutation $\Delta\Delta G$ values are only moderate, lower than 2.0 kcal/mol. Thus, our prediction overestimated their importance in this case. We noted that these two residues locate at both ends of the peptide ligand, which are exposed to the solvent. The solvent effect may compensate the loss in binding free energy once these salt bridges are broken due to residue mutation. Thus, the position of a residue also affects its probability of being a hot spot residue. Unfortunately, this type of consideration is not included in our current method.

Another example is given in Figure 3B, which shows the binding interface between Bcl-x$_L$ and the Bak-BH3 peptide. Here, Asp583 is the only false negative in our prediction. This residue is fully exposed to the solvent without forming any obvious interaction with Bcl-x$_L$ in this complex structure. Thus, it was not considered by our method as a residue involved in binding and has a zero hot spot index. However, experimental measurements indicate that mutating this residue to alanine resulted in a $\Delta\Delta G$ value of 2.86 kcal/mol, corresponding to an over 100-fold decrease in binding affinity. We assume that this residue affects the binding between Bcl-x$_L$ and Bak indirectly by stabilizing the helical structure of the Bak-BH3 peptide. Our current method cannot take into account this kind of effect either.

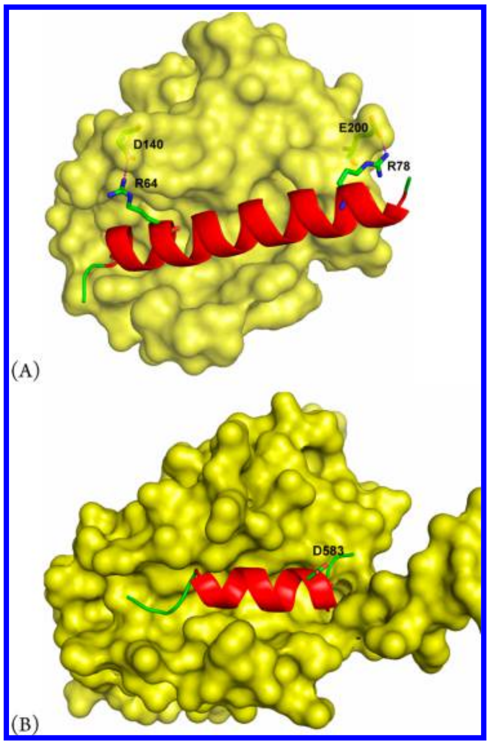


Figure 3. Illustration of two protein−protein complexes where our method made a wrong prediction of certain hot spots on the binding interface. (A) The Bcl-2/Bax complex (PDB entry: 2XA0), where Arg64 and Arg78 form salt bridges with Bcl-2 but they are not "hot spots" according to experimental data. (B) The Bcl-x_L/Bak complex (PDB entry: 1BXL), where Asp583 does not interact directly with Bcl-x_L but is a hot spot according to experimental data.

3.3. On Characteristic Interaction Patterns.

A central assumption in our method is that there are characteristic interaction patterns (CIPs) on protein−protein binding interface, which are responsible for preserving the biological functions of relevant protein molecules. As mentioned in the Introduction section, the concept of CIP resembles the concept of pharmacophore, which refers to a set of features in the structure of a drug molecule that are indispensable for reserving its biological activities. Those interaction patterns that have preference factors are logically the suitable candidates to be CIPs, yet a quantitative indicator is needed to measure the conservation degree of an interaction pattern across different protein−protein binding interfaces. For this purpose, the conservation score (eq 3) was introduced, which combines preference factor and geometry as well as composition of interaction pattern. The rationale for also including the latter two features is discussed below.

By our method, an interaction pattern is composed of four particular residues. The same set of residues, however, may appear in different geometrical arrangements, i.e. a tetrahedron in different geometries, on protein−protein binding interfaces. In our study, preference factors of all interaction patterns were derived by considering only their compositions but not their shapes. However, in order to interpret the relationships between CIPs on protein−protein binding interfaces and the biological functions of the relevant protein molecules, the geometrical arrangement of an interaction pattern should be taken into account. For example, several protein−protein binding interfaces sharing the same interaction pattern "D-YYR" are shown in Figure 4. This particular interaction pattern on binding interface 1ALL-AB and 1B8D-AB has good geometrical similarity, where the RMSD value between them is 0.086 Å. In fact, both complexes are involved in photosynthesis. In contrast, the same interaction pattern on binding interface 3DLQ-IR (Figure 4C) has a different geometrical arrangement, where the RMSD value to this pattern on binding interface 1ALL-AB is 1.347 Å. This complex is formed by cytokine and its receptor, which functions in intercellular communications. This example demonstrates the subtle role of the shape of CIPs in protein−protein interactions.

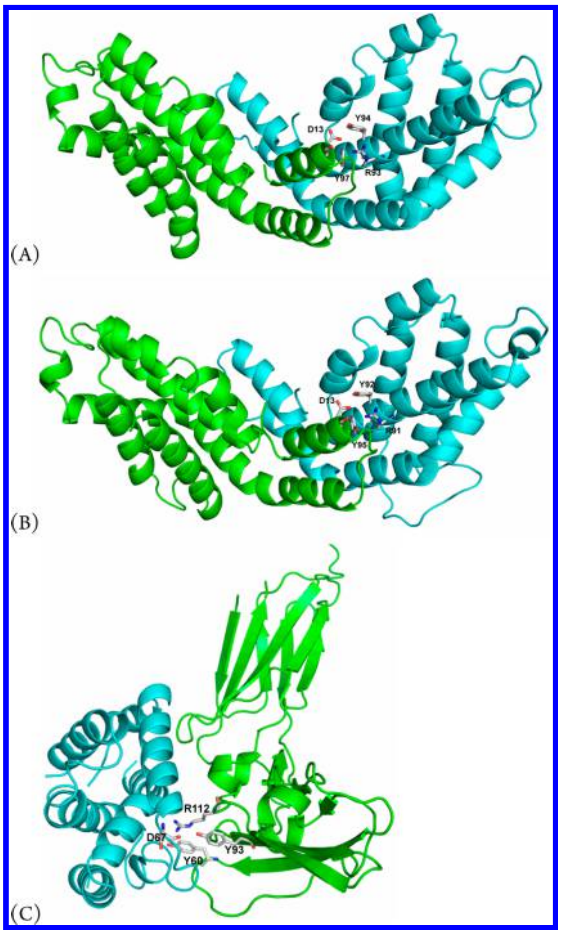


Figure 4. An example of common interaction patterns (D-YYR) found on different protein−protein binding interfaces: (A) PDB entry 1ALL; (B) PDB entry 1B8D; (C) PDB entry 3DLQ. The RMSD values of overlapping this interaction pattern between (A)/(B) and (A)/(C) are 0.086 Å and 1.347 Å, respectively.

Amino acid residues are classified into nine degenerated categories in our method to reduce the total types of interactions patterns while still reserving the essential chemical features of those residues. However, even the residues in the same category may play different roles. Thus, two identical interaction patterns with a higher level of residue degeneracy will tend to be more distant on the evolutionary tree. The residue similarity between two interaction patterns is measured by the Tanimoto coefficient, i.e. the number of identical residues divided by the total number of residues in two patterns (i.e., eight) minus the number of identical residues. For example, the Tanimoto similarity between the interaction patterns "S-ACQ" on the binding interface 1I51-AB and the binding interface 3EDQ-AB is 100% (Figure 5A and 5B). In fact, the complex in the former case is from caspase-7, and the

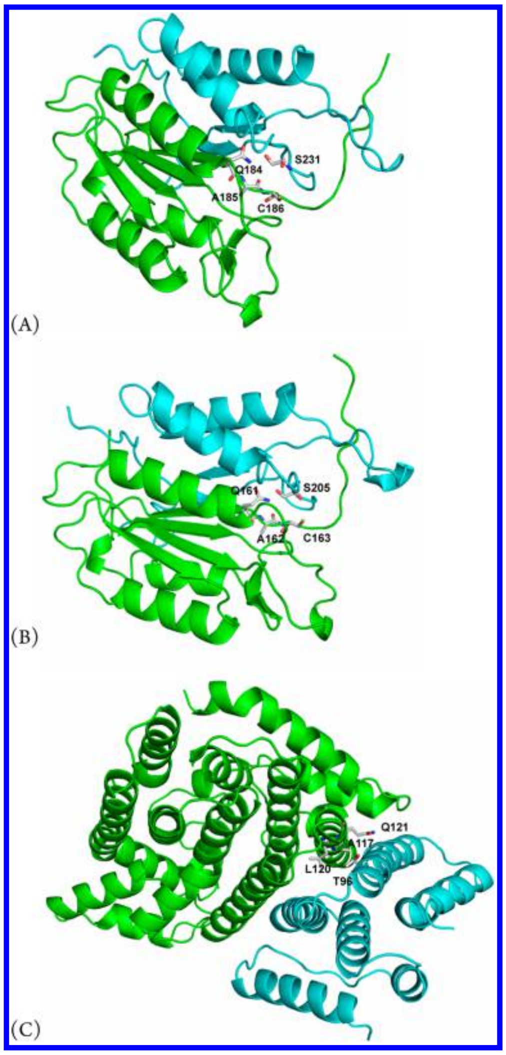


Figure 5. An example of similar interaction patterns observed on different protein−protein binding interfaces (A) PDB entry 1I51; (B) PDB entry 3EDQ; (C) PDB entry 3AQB. The Tanimoto similarity indices between the interaction patterns between (A)/(B) and (A)/(C) are 1.0 and 0.33, respectively. The complexes in entries 1I51 and 3EDQ are formed by caspases, which are apoptosis executors; whereas the complex in 3AQB is formed by a hexaprenyl diphosphate synthase.

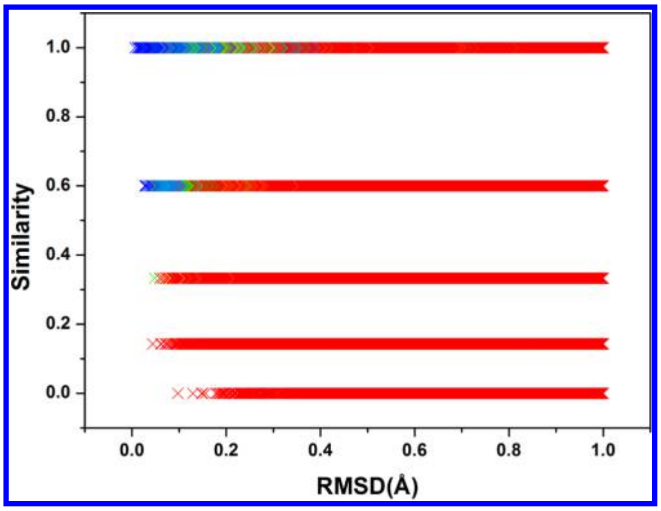


Figure 6. Distribution of the conservation scores for identical or similar interaction pattern pairs derived from protein−protein binding interfaces in Set-90. Only the interaction patterns with absolute RMSD < 2 Å are considered in this figure. The x axis is the geometrical similarity in normalized RMSD values. The y axis is the chemical similarity in Tanimoto similarity coefficients. The color scheme for indicating different levels of conservatism scores is blue (score ≥100), light blue (50 ≤ score <100), green (30 ≤ score <50), yellow (20 ≤ score <30), and red (score <20).

one in the latter case is from caspase-3, both of which belong to the caspase family. When the interaction pattern changes to "T-ALQ" on the interface 3AQB-AB (Figure 5C), the Tanimoto similarity drops to 33.3%. In fact, this complex is formed by a different protein, i.e. hexaprenyl diphosphate synthase. Therefore, preference factors should be supplemented by considering both geometrical and chemical similarities between interaction patterns to more faithfully reflect their significance on protein−protein binding interfaces.

A conservation score can be computed with eq 3 for each pair of interaction patterns on two protein−protein binding interfaces. Distribution of this score over the entire training set is given in Figure 6, where the x axis is normalized RMSD values and the y axis is Tanimoto similarity between two interaction patterns. It can be seen in this figure that interaction patterns with higher scores exist at the upper left region with low RMSD values (<0.2 Å) and relatively high Tanimoto similarity (60−100%). Only a few cases with moderate conservation scores exist at the upper right region with large RMSD values (0.2−0.5 Å) and high Tanimoto similarity (∼100%). In most cases of large RMSD values or small Tanimoto similarities, the conservation scores are low.

According to the distribution of conservation scores, we selected three cutoffs, i.e. 20, 30, and 50, to group protein−protein binding interfaces in our data set into clusters. The final clustering results on Set-50, Set-70, and Set-90 under three cutoffs are summarized in Tables S5−S7 in the Supporting Information. Since there can be multiple pairs of interaction patterns between two binding interfaces, the one with the highest conservation score was selected as the representative one. Then, we examined whether the protein−protein complexes in the same cluster belong to same protein family or have identical/similar biological functions by checking the annotations available on the PDB Web site. If so, it was counted as a correct cluster. The final results are summarized in Table 4. One can see that when a lower cutoff is applied to clustering, i.e. a lower level of conservation is required within each cluster, the total number of clusters will increase, and the rate of correct clustering will decrease. The same trend is observed on all three data sets. Here, the cutoff value of 30 seems to be appropriate for collecting protein−protein binding

Table 4. Correct Clustering of Protein−Protein Complexes in Terms of Their Biological Functions in Different Scenarios

	cutoffs of conservation scores used in clustering		
data sets	≥50	≥30	≥20
Set-50	96.55% (28/29)[a]	91.30% (42/46)	65.79% (50/76)
Set-70	95.92% (47/49)	92.11% (70/76)	65.69% (67/102)
Set-90	96.97% (64/66)	86.32% (82/95)	62.60% (77/123)

[a] The numbers outside brackets are clustering accuracy, whereas the numbers inside brackets are the ratio between the correct classified clusters and the total number of clusters.

interfaces with common CIPs, since it provides a good compromise between the accuracy in clustering and the total number of clusters.

We also adopted the evolutionary tree method to investigate the biological functions of the protein–protein complexes in each cluster. The graphical evolutionary tree for the protein–protein complexes which have common CIPs with conservation score >50 on their binding interfaces is given in Figure 7, which was generated by using the PHYLIP software[36] and then sketched with the TreeView program (version 1.6.6, http://taxonomy.zoology.gla.ac.uk/rod/treeview.html). This figure contains a total of 70 protein–protein binding interfaces belonging to 29 clusters from Set-50. Each protein–protein binding interface has two peptide chains, which are labeled after its PDB code with a suffix of "1" or "2", e.g. 1OZ7-1 and 1OZ7-2. On this evolutionary tree, the main branches usually have two or more sub-branches. Sub-branches stemming from the same main branch are in principle related closer to each other in evolution. Our examination of the sub-branches on each main branch reveals that they are all members in the same cluster with identical or similar functions. Take 1O94 and 3CLT for example, the two peptide chains in these two complexes locate on two sub-branches on the same main branch. which are both electron transfer flavoproteins. This evolutionary tree analysis provides additional support to our conclusions drawn based on conservation score.

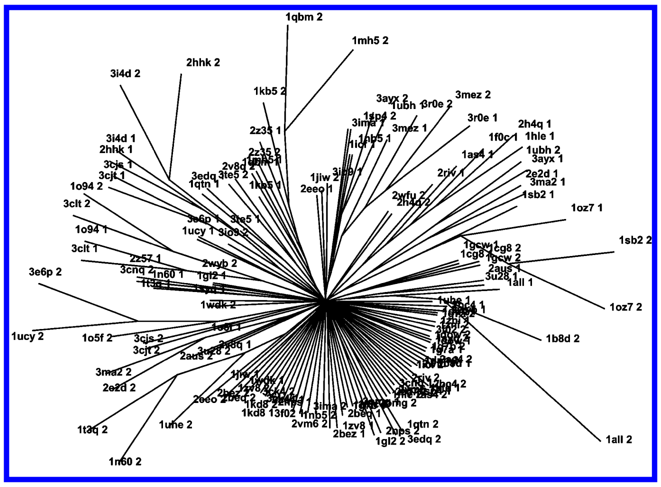


Figure 7. The evolutionary tree of the peptide chains from a total of 70 protein–protein complexes. Each branch denotes for one peptide chain involved in a certain protein–protein binding interface. Branches are clustered by a conservation score cutoff of 50.

We also compared the performance of our interaction pattern alignment with the ProBiS algorithm.[37] ProBiS aligns and superimposes complete protein surfaces, surface motifs, or protein binding sites. In this method, proteins are represented by defined functional groups similar to Mintz's approach. The comparison of binding sites is carried out through a maximal clique algorithm. ProBiS enables pairwise alignments as well as fast database searches for similar proteins or binding sites. Similar to the purpose of our analysis, ProBiS aims at detecting the functional relationship between protein molecules with similar binding sites. Here, the protein–protein interface clusters of Set-90 obtained by a conservation score of 50 were used as the test set since all members in these clusters have highly conserved CIPs on their binding interfaces and perform similar biological functions. The pairwise local structural alignment was then applied to those protein–protein interface clusters using the online ProBiS server at http://probis.cmm.ki.si/. In each cluster, Z-Scores between the first member and the remaining members were calculated (see Table S8 in the Supporting Information), where a higher Z-Score indicates a higher level of similarity. According to the default criterion adopted by ProBiS, a Z-Score above 2.0 indicates a similar structural alignment of two protein molecules. As a matter of fact, all of the computed Z-Scores are above 2.0. Some of them are even higher than 4.0. This observation suggests that the protein–protein complexes in each cluster are highly similar to each other in terms of structure. Thus, we conclude that both methods are able to correlate structural similarities with biological functions although they take different approaches.

4. CONCLUSIONS

We have developed a new method for mining the key factors in protein–protein interactions. Our method is distinctively different from those developed by other research groups, which are based on either residue pairs or functional groups. The central idea is to identify the so-called characteristic interaction patterns (CIPs). Each interaction pattern is composed of four contacting residues and therefore represents a microenvironment on a protein–protein binding interface. Our statistical survey on a nonredundant set of 1,220 protein–protein binding interfaces indicates that CIPs indeed have significantly higher occurrences than random combinations of four residues on protein–protein binding interfaces. Moreover, CIPs are found among protein–protein complexes of relatively high sequence similarity as well as those of relatively low sequence similarity.

The preference factor is used in our study to characterize the importance of an interaction pattern in the binding of two protein partners. A method for predicting hot spot residues is developed based on this quantity and was tested on two data sets. As compared to the predictions made by the Robetta method and the TSVM method, our method produced comparable recall scores but marginally poorer precision scores, which could be attributed to false positives. Unlike the other two methods, the performance of our method is basically independent of the choice of the training set. When it was applied to analyze the complexes formed between Bcl-2 family proteins, the precision score of our predictions was 0.60, and the recall score was 0.86. This level of accuracy is already meaningful for practical applications. Another indicator introduced in our study is the conservation score, which is used to group the protein–protein interfaces sharing common interaction patterns into clusters. Our statistical survey suggests that if protein–protein complexes share common CIPs, and if these CIPs have high conservation score, they are usually associated with identical or similar biological functions. This conclusion is also validated by the results derived from the evolutionary tree analysis conducted on our data set. Therefore, analysis of CIPs and their conservation scores provides an attractive approach for clustering protein molecules especially when the primary concern is their biological functions.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

A complete list of the protein–protein binding interfaces in our data set (Table S1), statistical results of preference factors

(Table S2), hot spot prediction results on the training set and the test set (Tables S3 and S4), the clustering results of protein–protein interfaces with similar interaction patterns (Tables S5–S7), and comparison of our method with ProBiS for aligning residue patterns (Table S8). This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: wangrx@mail.sioc.ac.cn.

Present Address

[§]State Key Laboratory of Bioorganic and Natural Products Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, 345 Lingling Road, Shanghai 200032, People's Republic of China.

Author Contributions

The manuscript was written through the contributions of all authors. In particular, Dr. Yan Li designed the methods, conducted most of the research, analyzed the results, and drafted the manuscript. Zhihai Liu and Chengke Li contributed substantially to data set compilation, method development, and results analysis. Prof. Renxiao Wang planned this project, supervised method development and result analysis, and also finished the manuscript. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work is supported by a Young Investigator Grant to Dr. Yan Li by the Chinese National Natural Science Foundation (Grant No. 21102168). The authors are also grateful for the financial support from the Chinese National Natural Science Foundation (Grants No. 81172984, 21072213, 21002117, 21102165, and 20921091), the Chinese Ministry of Science and Technology (863 High-Tech Grant No. 2012AA020308), and the Chinese Academy of Sciences.

■ REFERENCES

(1) Meszaros, B.; Tompa, P.; Simon, I.; Dosztanyi, Z. Molecular principles of the interactions of disordered proteins. *J. Mol. Biol.* **2007**, *372*, 549–561.

(2) Headd, J. J.; Ban, Y. E.; Brown, P.; Edelsbrunner, H.; Vaidya, M.; Rudolph, J. Protein-protein interfaces: properties, preferences, and projections. *J. Proteome Res.* **2007**, *6*, 2576–2586.

(3) Franzosa, E. A.; Xia, Y. Structural principles within the human-virus protein-protein interaction network. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 10538–10543.

(4) Jones, S.; Thornton, J. M. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 13–20.

(5) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **2007**, *25*, 71–75.

(6) Yan, C.; Wu, F.; Jernigan, R. L.; Dobbs, D.; Honavar, V. Characterization of protein-protein interfaces. *Protein J.* **2008**, *27*, 59–70.

(7) Ofran, Y.; Rost, B. Analysis of six types of protein-protein interfaces. *J. Mol. Biol.* **2003**, *325*, 377–387.

(8) Morrow, J. K.; Zhang, S. Computational prediction of protein hot spot residues. *Curr. Pharm. Des.* **2012**, *18*, 1255–1265.

(9) Grosdidier, S.; Fernandez-Recio, J. Protein-protein docking and hot-spot prediction for drug discovery. *Curr. Pharm. Des.* **2012**, *18*, 4607–4618.

(10) Tuncbag, N.; Gursoy, A.; Keskin, O. Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics* **2009**, *25*, 1513–1520.

(11) Fischer, T. B.; Arunachalam, K. V.; Bailey, D.; Mangual, V.; Bakhru, S.; Russo, R.; Huang, D.; Paczkowski, M.; Lalchandani, V.; Ramachandra, C.; Ellison, B.; Galer, S.; Shapley, J.; Fuentes, E.; Tsai, J. The binding interface database (BID): A compilation of amino acid hot spots in protein interfaces. *Bioinformatics* **2003**, *19*, 1453–1454.

(12) Kozakov, D.; Hall, D. R.; Chuang, G. Y.; Cencic, R.; Brenke, R.; Grove, L. E.; Beglov, D.; Pelletier, J.; Whitty, A.; Vajda, S. Structural conservation of druggable hot spots in protein-protein interfaces. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 13528–13533.

(13) Brenke, R.; Kozakov, D.; Chuang, G.-Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* **2009**, *25*, 621–627.

(14) Kim, D. E.; Chivian, D.; Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* **2004**, *32*, W526–W531.

(15) Tuncbag, N.; Keskin, O.; Gursoy, A. HotPrint: hot spot prediction server for protein interfaces. *Nucleic Acids Res.* **2010**, *38*, W402–W406.

(16) Darnell, S. J.; LeGault, L.; Mitchell, J. C. KFC server: interactive forecasting of protein interaction hot spots. *Nucleic Acids Res.* **2008**, *36*, W265–W269.

(17) Guharoy, M.; Pal, A.; Dasgupta, M.; Chakrabarti, P. PRICE (Protein Interface Conservation and Energetics): a server for the analysis of protein-protein interfaces. *J. Struct. Funct. Genomics* **2011**, *12*, 33–41.

(18) Mora, J. S.; Assi, S. A.; Fernandez-Fuentes, N. Presaging critical residues in protein interfaces-Web Server (PCRPi-W): a web server to chart hot spots in protein interfaces. *PLoS One* **2010**, *5*, e12352.

(19) Mintz, S.; Shulman-Peleg, A.; Wolfson, H. J.; Nussinov, R. Generation and analysis of a protein-protein interface data set with similar chemical and spatial patterns of interactions. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 6–20.

(20) Tsai, C. J.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. A data set of protein-protein interfaces generated with a sequence-order-independent comparison technique. *J. Mol. Biol.* **1996**, *260*, 604–620.

(21) Shulman-Peleg, A.; Mintz, S.; Nussinov, R.; Wolfsin, H. J. Protein-protein interfaces: Recognition of similar spatial and chemical organizations. In *Workshop on Algorithms in Bioinformatics*; Lecture Notes in Computer Science, Jonassen, I., Kim, J., Eds.; Springer Verlag: Bergen, Norway, 2004; Vol. *3240*, 194–205.

(22) Shulman-Peleg, A.; Shatsky, M.; Nussinov, R.; Wolfson, H. J. Spatial chemical conservation of hot spot interactions in protein-protein complexes. *BMC Biol.* **2007**, *5*, 43.

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

(24) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.

(25) Lise, S.; Archambeau, C.; Pontil, M.; Jones, D. T. Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods. *BMC Bioinf.* **2009**, *10*, 365.

(26) Kortemme, T.; Baker, D. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14116–14121.

(27) Huang, Z. The chemical biology of apoptosis. Exploring protein-protein interactions and the life and death of cells with small molecules. *Chem. Biol.* **2002**, *9*, 1059–1072.

(28) van Delft, M. F.; Huang, D. C. S. How the Bcl-2 family of proteins interact to regulate apoptosis. *Cell Res.* **2006**, *16*, 203–213.

(29) Lessene, G.; Czabotar, P. E.; Colman, P. M. BCL-2 family antagonists for cancer therapy. *Nat. Rev. Drug Discovery* **2008**, *7*, 989–1000.

(30) Hardwick, J. M.; Chen, Y. B.; Jonas, E. A. Multipolar functions of BCL-2 proteins link energetic to apoptosis. *Trends Cell Biol.* **2012**, *22*, 318−328.

(31) Sattler, M.; Liang, H.; Nettesheim, D.; Meadows, R. P.; Harlan, J. E.; Eberstadt, M.; Yoon, H. S.; Shuker, S. B.; Chang, B. S.; Minn, A. J.; Thompson, C. B.; Fesik, S. W. Structure of Bcl-x_L-Bak peptide complex: recognition between regulators of apoptosis. *Science* **1997**, *275*, 983−986.

(32) Petros, A. M.; Nettesheim, D. G.; Wang, Y.; Olejniczak, E. T.; Meadows, R. P.; Mack, J.; Swift, K.; Matayoshi, E. D.; Zhang, H.; Thompson, C. B.; Fesik, S. W. Rationale for Bcl-xL/Bad peptide complex formation from structure, mutagenesis, and biophysical studies. *Protein Sci.* **2000**, *9*, 2528−2534.

(33) Ku, B.; Liang, C.; Jung, J. U.; Oh, B.-H. Evidence that inhibition of BAX activation by BCL-2 involves its tight and preferential interaction with the BH3 domain of BAX. *Cell Res.* **2011**, *21*, 627−641.

(34) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*; University of California: San Francisco, 2006.

(35) Kelley, R. F.; O'Connell, M. P. Thermodynamic analysis of an antibody functional epitope. *Biochemistry* **1993**, *32*, 6828−6835.

(36) Felsenstein, J. *PHYLIP (Phylogeny Inference Package)* version 3.6; University of Washington: USA, 2005.

(37) Konc, J.; Janezic, D. ProBiS-2012: web server and web service for detection of structurally similar binding sites in proteins. *Nucleic Acids Res.* **2012**, *40*, W214−W221.