PROTEINS: Structure, Function, and Bioinformatics 64:68–78 (2006)

A Combinatorial Score to Distinguish Biological and Nonbiological Protein–Protein Interfaces

Shiyong Liu, Qingliang Li, and Luhua Lai*
State Key Laboratory for Structural Chemistry of Unstable and Stable Species, College of Chemistry and Molecular Engineering, Peking University, Beijing, China
Center for Theoretical Biology, Peking University, Beijing, China

ABSTRACT **With the large amount of protein–protein complex structural data available, to understand the key features governing the specificity of protein–protein recognition and to define a suitable scoring function for protein–protein interaction predictions, we have analyzed the protein interfaces from geometric and energetic points of view. Atom-based potential of mean force (PMFScore), packing density, contact size, and geometric complementarity are calculated for crystal contacts in 74 homodimers and 91 monomers, which include real biological interactions in dimers and nonbiological contacts in monomers and dimers. Simple cutoffs were developed for single and combinatorial parameters to distinguish biological and nonbiological contacts. The results show that PMFScore is a better discriminator between biological and nonbiological interfaces comparable in size. The combination of PMFScore and contact size is the most powerful pairwise discriminator. A combinatorial score (CFPScore) based on the four parameters was developed, which gives the success rate of the homodimer discrimination of 96.6% and error rate of the monomer discrimination of 6.0% and 19.8% according to Valdar's and our definition, respectively. Compared with other statistical learning models, the cutoffs for the four parameters and their combinations are directly based on physical models, simple, and can be easily applied to protein–protein interface analysis and docking studies. Proteins 2006;64:68–78.**

Key words: potential of mean force; contact size; packing density; combinatorial score; biological contact; protein–protein recognition

INTRODUCTION

Protein–protein recognition occurs in all kinds of life, which plays an important role in many biological processes such as signal transduction, gene expression control, enzyme inhibition, and antibody-antigen recognition. Protein interactions are widely studied by X-ray crystallography and computational and biochemical methods. In the past few years, many experimental techniques[1–4] and computational methods[5–7] have been used to study the functional networks of proteins in cells on the sequence level. Recently, the interactions between complexes were used to construct a structure-based network of molecular machines in the cell.[8] But the high false-positive problem for both experimental and computational methods became a bottleneck of building reliable protein–protein interaction networks.[9,10]

With the growing collections of protein three-dimensional (3D) coordinates deposited in Protein Data Bank (PDB),[11] high-throughput protein–protein docking may be used to build reliable protein–protein interaction networks and design novel functional receptors or inhibitors.[12] In fact, the computational protein–protein docking problem[13,14] is far from being solved, which is mainly due to our limited knowledge about the protein–protein recognition. Elucidating the principles governing protein–protein interactions at the atomic level will enrich our knowledge of protein interactions and improve the docking method.

We need to understand the fundamental questions about the physical chemistry of noncovalent protein–protein interactions. The known protein–protein complexes can be termed as biological complex, because they are known to associate in solution. Most crystal contacts are artifacts of crystallization that would not occur in solution, which are termed as nonbiological contacts. It is interesting to know what the main differences between biological and nonbiological contacts are. Physical and chemical properties of protein–protein interfaces have been analyzed by a number of research groups using protein complex structures from the PDB. The properties such as hydrophobicity, amino acid composition, hydrogen bonding, contact size, sequence conservation, shape, atomic

Abbreviations: PDB, Protein Data Bank; PMFScore, potential of mean force score; SCScore, shape complementarity score; CTPScore, combinatorial three-parameter score; CFPScore, combinatorial four-parameter score; ASA, accessible surface area.

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: Ministry of Science and Technology of China; Grant sponsor: NSFC; Grant numbers: 90403001, 30490245, and 20228306.

*Correspondence to: Luhua Lai, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China. E-mail: lhlai@pku.edu.cn

Received 4 October 2005; Revised 22 December 2005; Accepted 24 December 2005

Published online 4 April 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20954

packing density, and atomic contact vectors of the protein–protein interfaces have been studied.[15–22]

The previous studies show that biological contacts tend to be more specific and larger than nonbiological ones. By using these parameters as a checking procedure, it was observed that they could not differentiate between biological and nonbiological contacts unambiguously. Bahadur et al.[22] combined the nonpolar interface area and the fraction of buried interface atoms and assigned 88% of the biological contacts of the homodimers and 77% of the nonbiological contacts of the monomers in their data set correctly. At the same time, they reported that unlike the specific interface, the large crystal-packing interfaces are poorly packed. A potential based on just two atom types (hydrophobic and hydrophilic) has been used to identify native-like complexes by their favorable potential energies.[23] The hydrophobic potential was shown to perform better. Biological interfaces contain many specific interactions, including van der Waals attraction and repulsion, hydrogen-bond interactions, electrostatic interactions, etc. The average energy contribution of H-bond is 47%; and that of hydrophobic interaction is 50%.[24]

To further analyze the contribution of different interactions to protein–protein complex and to formulate a combined criteria to differentiate biological and nonbiological protein contacts, we have used atomic based potential mean force score,[25] contact size, packing density,[12] and shape complementarity[26] of the protein–protein interface in the study. The combination of packing density and potential of mean force has never been used to discriminate between nonbiological and biological interfaces. A combinatorial score based on the four parameters was developed and successfully applied to differentiate biological and nonbiological protein–protein interfaces.

MATERIALS AND METHODS

Data Collection

The data set of Ponstingl et al.[27] was used to provide a starting point for further analysis; 76 homodimers and 92 monomers were downloaded from http://www.ebi.ac.uk/thornton-srv/software/quasiprox/. 1kba and 1vlb were excluded because 1kba is a homododecameric complex, and the biological interface of 1vlb is not clearly defined (entry 1vlb replaces 1hlr, 1alo. The asymmetric unit for 1vlb contains one copy of the macromolecule. "REMARK 350" gave an identity matrix for generating the biomolecule. So the biological interface of 1vlb is not clear). The PITA program written by Ponstingl et al.[28] was used to generate all crystal contacts for each structure. Because the largest contact size of interface of cytochrome c3 from desulfovibrio desulfuricans Norway (PDB 2cy3) is 7, it was obviously monomer and excluded from our monomer data set. Thus, 74 homodimers and 91 monomers are kept. If there are several identical interfaces, only one is kept. Among the 74 homodimers, 1hjr (A:C, B:D), 4kbp (A:E, B:C), and 9wga (A:C, B:D) have two biological interfaces. The crystal contacts with contact size larger than 10 (about B = 567 Å^2) are kept and used in our analysis. Ponstingl et al.[27] only selected the crystal contacts with the largest contact area and made an arbitrary choice when the maximum contact size was not unique. Mintseris and Weng[21] used the same criteria as Ponstingl.

But for some dimers, the largest interfaces are not the biological one. For example, the biological dimers of farnesyl pyrophosphate synthetase (1uby) and carboxylesterase from pseudomonas fluorescens (1auo) correspond to the second largest interfaces. For the crystal contacts of monomers, sometimes several contacts are comparable in size, and it is more reasonable to include more monomer interfaces of comparable size in the study. This is particularly important if we want to derive general criteria to discriminate real protein–protein interactions and for docking studies.

The biological interfaces of the homodimers were further confirmed by literature reading. In total, 296 interfaces from 74 homodimers and 465 interfaces from 91 monomers, which include 77 true biological interfaces, are used in this study. In all the cases, the dimer interfaces are the crystal contacts with the largest interface but 1uby and 1auo (see Supplemental Material).

The docking decoy set I consists of 16 decoy sets downloaded from the Sternberg group's Web site (http://www.bmm.icnet.uk/docking/). The docking decoy set II consists of 42 decoy sets downloaded from Weng group's Web site (http://zlab.bu.edu/zdock/decoys.shtml).

Definition of Protein–Protein Interface

The interface is defined as the set of atoms on a protomer that each lose at least 0.1 Å^2 of accessible surface area (ASA) on binding with a partner and with <15 Å^2 of solvent ASA.[12] Atom and residue solvent ASA are calculated by using the program NACCESS[29] with a 1.4 Å probe radius. The interface area is measured by comparing the solvent ASA of the complex to that of its components[15]:

$$B = A_A + A_B - A_{AB} \quad (1)$$

where A_{AB} is the solvent ASA of the complex, and A_A and A_B are solvent ASA of both protomers, respectively. B represents the buried surface area of the two component proteins in contact.

PMFScore

The PMFScore is a statistical potential developed by Jiang et al.[25] to estimate binding free energy for protein–protein interactions. The following four atom types were used in the PMFScore calculation: hydrogen bond donor, hydrogen bond acceptor, both donor and acceptor, and neutral atom (neither donor nor acceptor). The details of the specific definitions have been described in Jiang et al.[25] The PMFScore is defined as:

$$\mathrm{PMFScore} = \sum_{ij} \Delta A_{ij}(r) \times \Delta_{ij} \quad (2)$$

where

$$A_{ij} = -kT \ln[f_{ij}(r) / Z_{ij}] \quad (3)$$

where k is the Boltzmann constant and T is the absolute temperature; $f_{ij}(r)$ is a frequency of ij contacts occurring at distance r.

$$A_{ij} - \text{reference energy} \equiv \Delta A_{ij}(r) = kT \ln[1 + m_{ij}\sigma] - kT \ln\left[1 + m_{ij}\sigma \frac{g_{ij}(r)}{f(r)}\right]$$

where m_{ij} is the total number of contacts between types i and j, σ is set to 0.02, $g_{ij}(r)$ is the distribution of these contacts occurring at distance r, and f(r) is the distribution of all contacts for all types at distance r. And

$$\begin{cases} \Delta_{ij} = 0 \text{ for } r_{ij} > r_{\text{cut-off}} \\ \Delta_{ij} = p_i \times p_j \text{ for } r_{ij} \le r_{\text{cut-off}} \end{cases} \quad (4)$$

where r_{cutoff} is the cutoff distance of atom-type pair interactions and is set to 8.0 Å, and p_i is a weighting coefficient from the atomic occupancy. According to Jiang's result, the correlation coefficient between PMFScore and experimental binding free energy $\Delta G_{\text{binding}}$ is about 0.75.

SCScore

This method of measuring shape complementarity by Fourier correlation has been widely used in protein–protein docking. Jackson et al.[26] released the docking program FTDock in 1998, which is available free-of-charge to academic and nonprofit researchers. FTDock was modified to output a SCScore for a given protein–protein interface. We assign a protein–protein complex structure as two parts: receptor A and ligand B. First, molecules A and B are discretized in a 3D N × N × N grid with every grid point (l,m,n = {1…N}) assigned a value:

$$f_{A_{l,m,n}} = \begin{cases} 1\text{:surface of molecule} \\ \rho\text{:core of molecule} \\ 0\text{:open space} \end{cases} \quad (5)$$

and

$$f_{B_{l,m,n}} = \begin{cases} 1\text{:inside molecule} \\ 0\text{:open space} \end{cases} \quad (6)$$

We kept ρ the same as that used in their docking study. Let's consider the protein complex AB, the translation vector of molecule B relative to A is p, q, r. To calculate the SCScore of AB, a correlation function of f_A and f_B is defined as:

$$\text{SCScore}_{r,s,l} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} f_{A_{l,m,n}} \times f_{B_{l+r,m+s,n+t}} \quad (7)$$

Thus, we can easily calculate SCScore from a protein complex structure according to the above equations. A high SCScore means that protein–protein interface of the complex has good surface complementarity.

Packing Density

The program for calculating packing density was a part of a protein functional grafting program, which was developed by Liang et al.[12] Packing densities were calculated according to Richards by using method B[30] A 2.8 Å solvent shell was added onto the protein surface. Packing density for each of interface atoms was calculated individually and averaged.

Definition of Accuracy and Error Rate

To assess the performance of the four parameters and their combinations, we use the following definitions by Valdar and Thornton[19]:

- p = number of correctly classified biological contacts
- n = number of correctly classified nonbiological contacts
- o = number of nonbiological contacts classified as biological (overpredictions)
- u = number of biological contacts classified as nonbiological (underpredictions)
- t = p + n + o + u
- accuracy = (p + n)/t × 100%
- monomer error rate = 100% − accuracy

To penalize many more nonbiological contacts in the homodimer data set, a ϕ-coefficient[19] is taken into consideration when measuring performance of predictors. The ϕ (Matthew's correlation coefficient) measures the correlation between observed and predicted results and ranges from −1 to +1, and the +1 means perfect correlation and an ideal prediction. The ϕ is calculated as:

$$\phi = \frac{pn - ou}{\sqrt{(p+o)(p+u)(n+o)(n+u)}} \quad (8)$$

Performance on the homodimers is assessed by using accuracy and ϕ. On the other hand, the error rate is used to measure performance of predictor on the monomers. Because of the large number of nonbiological contacts, Valdar's definition usually gave a smaller error rate. In fact, once a contact of monomer is predicted to be biological, this monomer will be misclassified as biological dimers. Thus, we defined a more strict error rate for the monomers as:

- f = number of monomers classified as biological dimers
- g = total number of monomers
- error rate = f/g ×100%.

RESULTS

Analysis of Interface by Four Independent Parameters

PMFScore

We calculated PMFScore (potential mean force score) of 296 interfaces from 74 homodimers and 465 interfaces from 91 monomers, which includes 77 true biological interfaces, and 684 nonbiological interfaces. The PMFScore distribution of nonbiological interfaces of homodimers is modeled with a Gaussian curve. And the value of −66551 is the three standard deviations from the mean. According to our calculation of PMFScore on the interfaces from homodimers, a contact is predicted to be biological if its PMFScore is lower than −66551, which is determined by the PMFScore distribution of nonbiological contacts from homodimers. The overall accuracy by this

predictor on the homodimers is 94.3% with a relatively low Φ value of 0.85. By using the same predictor, 35 contacts (19 monomers) among the 465 nonbiological contacts from 91 monomers are classified as biological. The error rate is 20.9% for the monomers according to our definition.

Contact size

The contact size is defined as the number of residues on the protein–protein interface including both sides of the two-component proteins. The contact size scales linearly with the interface area. Crystal interfaces from homodimers comprise one residue per 78.5 Å^2 (correlation coefficient $R^2 = 0.98$; intercept = -217.9 Å^2). The contact size 22 [about B = 1509 Å^2 (i.e., is 755 Å^2 per subunit)] is selected as a predictor according to the distribution of the dimer data set. A contact is predicted to be biological if its contact size is bigger than 22. Biological contacts were typically bigger than nonbiological contacts, but there is a region of overlap between them. This predictor gave an overall accuracy of 96.3% with a Φ value of 0.90 for the homodimers. When the same predictor was applied to the monomer data set, 52 nonbiological contacts (29 monomers) among the 465 contacts are classified as biological. The error rate is 31.9% for the monomers according to our definition. To identify biological contacts with a smaller contact size or select crystal-packing interfaces comparable in size to biological interfaces, it is obvious that more than one parameter is required. The combination of contact size and some other parameters will be necessary.

SCScore

The SCScore (shape complementarity score) is used to analyze protein–protein interface. The calculation protocol of SCScore used in this study is similar to that used in FTDock,[26] which follows closely the shape recognition algorithm based on the method of Katchalski-Katzir et al.[31] SCScore of the same homodimer and monomer data sets were calculated. Biological contacts are often more highly shaped complementary than nonbiological ones. By using the dimer data set as a training set, the contact is predicted to be biological if its SCScore is bigger than 108. This predictor gave an overall accuracy of 94.3% and Φ value of 0.85 for the homodimers. When the same predictor was applied to the monomer data set, 46 nonbiological contacts (28 monomers) among the 465 contacts are classified as biological. The error rate is 30.8% for the monomers according to our definition.

Packing density

A local atomic density (LD) index and a global density index (GD)[22] have been shown that these indices are larger at biological than at nonbiological interfaces. Our packing densities are calculated according to Richards using method B.[30] The distributions of packing densities show that biological contacts tend to be better packed than nonbiological contacts. Packing density of the biological contacts in the homodimer data set is between 0.659 and 0.783. On average, biological contacts have a packing density of 0.72, whereas nonbiological contacts have an average of 0.70 in the homodimer data set and 0.707 in the monomer data set. So, according to the packing density distribution, when the packing density of an interface is <0.60 or >0.84, it should be a nonbiological interface. This rule is very useful in filtering docking decoys.

Analysis of Interface by Pairwise Combination of the Four Parameters

Analysis of interface by PMFScore and contact size

The combination of cutoffs from PMFScore and contact size of homodimer data set, which is called ① ③ in Table I, gave an overall accuracy of 97.0% with a Φ value of 0.92 for the homodimers. The predictor named ① ③ in Table I has the form:

$$① ③ \equiv \text{Score13} \equiv \frac{\text{Contact size}}{22.0} + \frac{\text{PMFScore}}{(-66551.0)} - 2.0$$

and

$$\begin{cases} \text{Score13} \geq 0.0\text{:biological contact} \\ \text{Score13} < 0.0\text{:non-biological contact} \end{cases} \quad (9)$$

Thirty nonbiological contacts (19 monomers) among the 465 contacts are classified as biological, when the same predictor was applied to the monomer data set. The error rate of monomer classification is 20.9% by our definition and 6.4% by Valdar's definition, respectively.

Analysis of Interface by Other Pairwise Combination of the Four Parameters

PMFScore and contact size are the best pairwise combination (① ③ in Table I) for distinguishing biological and nonbiological contacts. Besides ① ③, the results of other pairwise combination of the four parameters were also listed in Table I. The form of other pairwise combination is similar to Eq. (9). The predictors related to PMFScore gave a better prediction (error rate about 20%) on monomer data set than others (error rate about 30%).

Analysis of Interface by a Combinatorial Score From Three Parameters

The performances of pairwise combinations from the four parameters (PMFScore, contact size, and packing density) have been listed above. Their accuracies (Table I) are about 94–97% for the homodimers, but the error rate is still 20–34% for the monomers. We then made a simple combination of the cutoffs of the four parameters: PMFScore (-66551.0), packing density (0.72), and contact size (22.0) of protein–protein interface as a new discriminator, which is expected to be better in distinguishing biological from nonbiological crystal contacts in 74 homodimers and 91 monomers. The following definition of the combinatorial three-parameter (CTPScore) score is derived from three cutoffs:

$$\text{CTPScore} \equiv \frac{\text{Contact size}}{22.0} + \frac{\text{PMFScore}}{(-66551.0)} + \frac{\text{Packing Density}}{0.72} - 3.0$$

and

TABLE I. Summary Statistics for All Predictors From Cutoffs

	Homodimers			Monomers		
Predictors[a]	p/n/o/u	Accuracy[b]	Value[c]	o/f	Error rate[d]	Error rate[e]
①	63/216/3/14	94.3%	0.85	35/19	20.9%	7.5%
②	39/140/79/38	60.5%	0.13	155/60	65.9%	33.3%
③	70/215/4/7	96.3%	0.90	52/29	31.9%	11.2%
④	68/211/8/9	94.3%	0.85	46/28	30.8%	9.9%
①②	65/216/3/12	94.9%	0.87	33/18	19.8%	7.1%
①③	69/218/1/8	97.0%	0.92	30/19	20.9%	6.4%
①④	65/215/4/12	94.6%	0.86	32/19	20.9%	6.9%
②③	71/215/4/6	96.6%	0.91	46/27	29.7%	9.9%
②④	67/210/9/10	93.6%	0.83	43/27	29.7%	9.2%
③④	70/214/5/7	96.0%	0.89	47/31	34.1%	10.1%
①②③	69/218/1/8	97.0%	0.92	31/19	20.9%	6.7%
②③④	72/214/5/5	96.6%	0.91	47/31	34.1%	10.1%
①③④	69/215/4/8	96.0%	0.89	28/18	19.8%	6.0%
①②④	67/215/4/10	95.3%	0.88	31/18	19.8%	6.7%
①②③④	71/215/4/6	96.6%	0.91	28/18	19.8%	6.0%
SVM0	74/218/1/3	98.6%	0.96	32/21	23.1%	6.9%
SVM1	69/218/1/8	97.0%	0.92	12/7	7.7%	2.6%

[a]①, ②, ③, ④ are PMFScore, Packing density. Contact size, and SCScore.
[b]Correctly classified contacts, see Materials and Methods.
[c]Measures the correlation between observed and predicted results.
[d]Error rate from our definition; see Materials and Methods.
[e]Error rate from Valdar's definition; see Materials and Methods.

$$\begin{cases} \text{CTPScore} \geq 0.0\text{:biological contact} \\ \text{CTPScore} < 0.0\text{:non-biological contact} \end{cases} \quad (10)$$

When the three cutoff parameters were combined as a new discriminator, the accuracy of discriminating homodimers reaches 97.0% with the Matthew's correlation coefficient of 0.92, and error rate of the monomers is lowered to 20.9%. The CTPScore here is the predictor: ① ② ③ in Table I, which outlined the results of all other CTPScores. From Table I, the predictor ① ② ③ has the best performance among the three-parameter score (CTPScore) on the homodimers and monomers if the accuracy Φ value and error rates are all taken into consideration.

Analysis of Interface by Combinatorial Score From Four Parameters

We have tried to use single, pairwise, and triplewise parameters. The results have been summarized in Table I. Then a simple combination of the cutoffs of the four parameters [PMFScore (−66551.0), packing density (0.72), SCScore (108.0), and contact size (22.0)] of protein–protein interface is used as a new discriminator. The following definition of the combinatorial four-parameter score (CFPScore) is derived from the four cutoffs:

$$\text{CFPScore} \equiv \frac{\text{Contact size}}{22.0} + \frac{\text{PMFScore}}{(-66551.0)} + \frac{\text{SCScore}}{108.0} + \frac{\text{Packing Density}}{0.72} - 4.0$$

and

$$\begin{cases} \text{CFPScore} \geq 0.0\text{:biological contact} \\ \text{CFPScore} < 0.0\text{:non-biological contact} \end{cases} \quad (11)$$

When the four cutoff parameters were combined as a new discriminator, the accuracy of discriminating homodimers reaches 96.6% with the Matthew's correlation coefficient of 0.91, and error rate of the monomers is lowered to 19.8% of our definition and 6.0% of Valdar's definition, respectively. The distributions of CFPScore for biological and nonbiological interfaces are shown in Figure 1. It is easy to discriminate between biological and nonbiological interfaces for homodimers using CFPScore as a single cutoff.

To vary the weights of the four individual scores, support vector machine (SVM) method[32] was used to determine the five parameters P1, P2, P3, P4, and P5 in the following equation.

$$\text{SVM} \equiv \frac{\text{Contact size}}{22.0} \times \text{P1} + \frac{\text{PMFScore}}{(-66551.0)} \times \text{P2} + \frac{\text{SCScore}}{108.0} \times \text{P3} + \frac{\text{Packing Density}}{0.72} \times \text{P4} - \text{P5}$$

$$\begin{cases} \text{SVM} \geq 0.0\text{:biological contact} \\ \text{SVM} < 0.0\text{:non-biological contact} \end{cases} \quad (12)$$

The first SVM model (SVM0) was trained on 296 interfaces from 74 homodimers. The accuracy of homodimer discrimination reaches 98.6% with the Matthew's correlation coefficient of 0.96, and error rate of the monomers is lowered to 23.1% of our definition and 6.9% of Valdar's definition, respectively. The second (SVM1) model was trained on 296 interfaces from 74 homodimers and 465 interfaces from 91 monomers, which includes 77 true biological interfaces, and 684 nonbiological interfaces. The

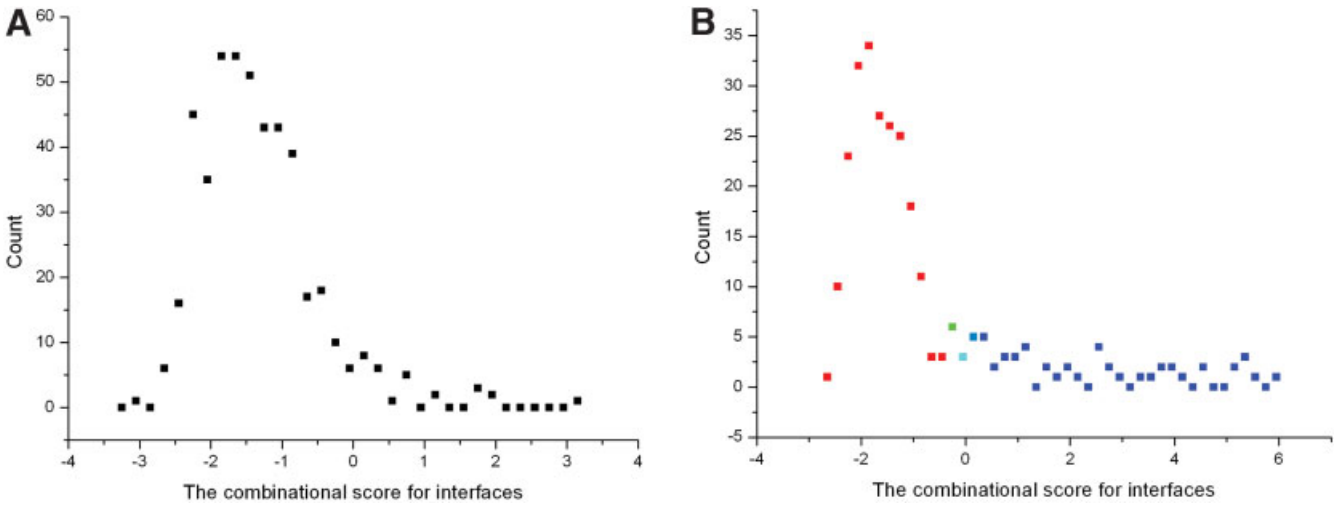

Fig. 1. Distribution of CFPScore for homodimers and monomers. The cutoff for CFPScore is zero. When the CFPScore of the interface is bigger than zero, it is predicted to be biological. **a:** Distribution of parameters for monomers. **b:** Distribution of parameters for homodimers red: nonbiological contacts; blue = biological contacts.

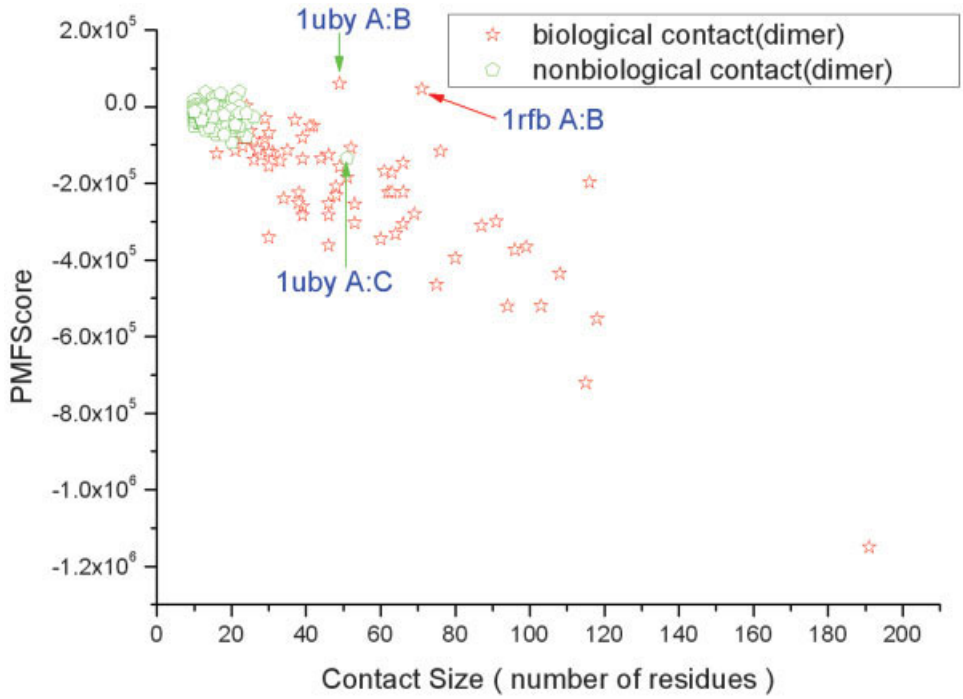

Fig. 2. PMFScore and contact size of biological and nonbiological interface of dimer. Contact size is plotted as the number of residues.

TABLE II. Parameters Used for the Two SVM Models

	P1	P2	P3	P4	P5	p/n/o/u
SVM0	4.56	2.17	0.22	15.61	22.38	74/218/1/3
SVM1	2.83	1.18	0.39	18.06	23.07	69/670/15/8

parameters are listed in Table II. It is apparent from Table I that SVM models improve slightly more than CFPScore to differentiate biological and nonbiological protein–protein interfaces for crystal structures. Both SVM0 and CFPScore have been trained on the interfaces from homodimers and give better predictions for dimer than monomers, whereas SVM1 is trained on the interfaces from homodimers and monomers and gives significant improvement for differentiating monomers. Because the purpose here is to derive a consensus score that can be applied to docking studies, we do not think it very useful to try to fit the crystal contact data with complicated machine learning methods. To choose a better score for protein–protein docking studies, two docking decoy sets were tested in the following section.

TABLE III. Ranking of the Native State and the Z-Score for the 16 Docking Decoy Sets

PDB ID	LLS[a]	DFIRE[b]	SVM0	SVM1	CFPScore
1avz	2	1/3.31[c]	5/1.45	14/0.45	5/1.61
1bgs	1	1/4.56	3/1.88	20/0.63	1/2.77
1brc	1	1/3.50	1/3.11	3/1.38	1/3.85
1fss	1	1/4.33	1/1.83	4/0.82	1/2.77
1ugh	1	1/4.85	2/1.87	12/0.69	1/3.10
1wql	1	1/5.29	19/0.20	33/−0.37	1/1.83
2pcc	1	1/2.86	2/1.50	5/0.73	7/1.12
2sic	1	1/4.29	1/2.66	1/1.80	1/3.58
1cgi	1	1/5.89	1/2.67	1/1.85	1/3.25
1dfj	4	1/4.11	3/0.41	14/−0.15	1/1.26
1ahw	3	1/3.80	1/2.92	1/2.25	1/4.14
1bvk	4	1/3.50	1/4.74	1/4.08	1/4.42
1dqj	4	1/4.85	2/1.07	19/−0.08	1/2.17
1mlc	3	1/4.04	12/0.58	50/−0.61	11/0.89
1wej	1	1/3.10	1/2.53	3/1.46	1/2.96
2kai	14	1/4.25	1/3.17	1/1.96	1/3.64
% Success[d]	9/16 (56%)	16/16 (100%)	8/16 (50%)	5/16 (30%)	13/16 (81%)

[a]The residue-specific all-atom knowledge-based potential by Lu, Lu, and Skolnick[33] derived from the interfacial structures of a dimer database.
[b]The DFIRE-based potential derived from a structure database of single-chain proteins.[34]
[c]The number in each cell indicates the rank of the native structure and the Z-score, respectively. (The Z-score was not reported in Lu, Lu, and Skolnick[33]).
[d]The overall success rate based on the first rank.

Application in Docking Decoy Sets

Native structure selection from 16 docking decoy sets

To test the ability of the combinatorial score (CFPScore, SVM0, and SVM1) in docking study, we used the docking decoy sets from the Sternberg group (http://www.bmm.icnet.uk/docking/). For each target, the data set includes 100 decoys. When the packing density of an interface is <0.60 or >0.84, it should not be a biological interface. This simple rule has been applied to the docking decoy set before filtering by our scores. Table III compares the results of the CFPScore, SVM0, and SVM1 with those of the all-atom knowledge-based Lu-Lu-Skolnick (LLS) potential[33] and DFIRE-SCM potential developed by Zhang et al.[34] The overall success rate based on the first-rank solutions shows that CFPScore is significantly better than SVM0, SVM1, and LLS, although DFIRE-SCM potential performed best with this data set. A z-score definition similar to Zhang et al.[34] was used for the comparison.

Reranking 42 docking decoy sets

We also used CFPScore to rerank the docking decoy sets used by RDOCK.[35] CFPScore is compared with RDOCK in Table IV. Rank1, Rank2, and root-mean-square deviation (RMSD) in Table IV are supplied by Li et al.[35] Rank1 is the rank of the best ranked hit by RDOCK; Rank2 is the original rank by RDOCK of the best ranked hit by CFPScore; Rank3 is the rank of the best ranked hit by CFPScore. Rank3 is calculated by CFPScore. To make a simple comparison, for the same test case, if the difference of the first rank (Rank1 and Rank3 in Table IV) is smaller than 10, they are considered to be no difference because the structures in this data set are densely populated. CFPScore improves the rank of the first hit for 7 test cases, worsens the rank for the 12 test cases, and does not change too much in the remaining 23 (54.8%) test cases. When the SVM model is compared with RDOCK, it worsens the rank slightly more than CFPScore (data not shown).

DISCUSSION

PMFScore Is More Significant Than Contact Size, Packing Density, and SCScore for Distinguishing Biological and Nonbiological Contacts

The PMFScore is more significant than contact size, packing density and SCScore for distinguishing biological and nonbiological contacts. For PMFScore, the best-separating cutoff was −66551.0 for the homodimers. With this cutoff, 19 of 91 monomers were misclassified as dimers and a relatively low Φ value of 0.85 with an accuracy of 94.3% for the homodimers. For contact size, the best-separating cutoff was 22.0 for the homodimers. With this cutoff, 29 of 91 monomers were misclassified as dimers and a relatively high Φ value of 0.90 with an accuracy of 96.3% for the homodimers. Mintseris and Weng[21] used atomic contact vectors to distinguish homodimers and crystal contacts on the same data set used by Ponstingl et al.[27] Using quadratic Fisher discriminator and kernel discriminator analysis, they reached a success rate of 93.0% for the leave-one-out cross-validation tests. Because our purpose is to search for general purpose scoring function from the homodimer and crystal contacts data set that can be applied in other studies, the PMFScore used here was derived from a data set of 191 heterodimer interfaces.[25] Only a simple cutoff was used here and no training was done. Our success rate was 94.3% for homodimers, which is comparable with Mintseris and Weng's result.

TABLE IV. Performance of the CFPScore on Docking Decoy Set by Li et al[35]

	RDock			CFPScore		
Complex[a]	Decoys[b]	Rank1[c]	RMSD[d]	Rank2[c]	Rank3[c]	RMSD[d]
1CGI	152	8	2.24	88	3	2.41
1CHO	137	1	1.28	12	7	0.91
1TGS	201	8	2.03	8	7	2.03
1BRC	115	3	2.41	3	5	2.41
1ACB	120	1	1.86	1	2	1.86
1MAH	106	1	0.91	1	10	0.91
1UGH	104	1	2.08	6	4	2.37
1DFJ	110	1	2.48	8	3	1.86
1FSS	105	42	1.52	108	48	1.70
1PPE[e]	359	1	0.69	11	1	0.76
1TAB[e]	121	10	0.76	40	10	1.05
1STF[e]	142	1	1.04	3	1	0.96
2TEC[e]	177	1	0.83	87	2	1.84
1WEJ	104	4	0.91	95	2	1.83
1AHW	128	1	1.61	36	5	1.35
1FBI[e]	101	53	2.15	53	43	2.15
1BOL[e]	115	1	1.18	1	10	1.18
1NCA[e]	150	8	0.83	9	1	2.46
1MEL[e]	151	1	1.37	7	7	1.47
1QFU[e]	110	29	0.95	29	19	0.95
1WQI	125	16	1.91	36	25	2.31
1IGC[e]	106	21	1.18	103	22	1.32
1SPB[e]	168	1	0.70	2	1	0.59
2KAI	103	141	2.36	507	18	2.47
1BRS	132	13	1.23	95	1	2.48
1JTG	159	13	1.54	115	1	2.26
1DQJ	101	952	2.45	952	64	2.45
1BVK	102	1314	1.64	1419	72	1.89
2JELe	157	301	1.7	744	13	1.71
1JHLe	115	41	0.88	1026	21	2.02
2PTC	102	2	1.12	2	74	1.12
2SIC	123	1	1.17	9	23	1.97
1CSE	103	1	1.17	4	58	0.91
1AVW	124	2	2.00	2	19	2.00
1UDI[e]	115	3	1.06	23	18	1.10
4HTC[e]	101	1	1.46	2	28	1.27
1MLC	103	2	1.65	841	70	2.40
1NMB[e]	106	1	1.11	6	26	1.10
2VIR[e]	103	80	1.19	264	93	1.34
1ATN[e]	101	1	0.80	1	82	0.80
2BTF[e]	111	1	0.95	274	15	1.33
1A0O[e]	102	11	2.46	11	80	2.46
Total[f]				23/42 (54.8%)		
Total[g]				7/42 (16.7%)		
Total[h]				12/42 (28.6%)		

[a]4-letter Protein Data Bank (PDB) code for the crystal complex of a test case.
[b]Each test case consists of the first 100 false positives (after RDOCK) and hits (Hits are defined as docked structures with interface C_a RMSD $\leq$ 2.5 Å from the crystal complex.)
[c]Rank1: Rank of the best ranked hit by RDOCK. Rank2: The initial rank (by RDOCK) of the best ranked hit by CFPScore. Rank3: Rank of the best ranked hit by CFPScore.
[d]Interface C_a RMSD for the best ranked hit.
[e]Unbound/bound test cases.
[f]$|\text{Rank1} - \text{Rank3}| \leq 10$.
[g]$\text{Rank1} - \text{Rank3} > 10$.
[h]$\text{Rank3} - \text{Rank1} > 10$.

Most biological interfaces of the 74 dimers but 1rfb (A:B) have a reasonable PMFScore compared with nonbiological interface. Figure 2 plots the distribution of PMFScore for the biological and nonbiological contacts in the homodimer data set. In Figure 2, 1uby (A:B, symmetry: 1/2-X, Y, 3/4-Z, 49 residues in contact) has a positive PMFScore. Farnesyl pyrophosphate synthetase (PDB 1uby) is active as a homodimer,[36] which was proposed to be related by a

crystallographic twofold axis.[37] The biological state given here for 1uby (A:B, symmetry: 1/2-X, Y, 3/4-Z, 49 residues in contact) should be biological according to the biological unit description in PDB by Tarshis et al.[37] However, after checking the annotations from PQS (the Protein Quaternary Structure) server,[38] 1uby (A:C, symmetry: -X, Y, -Z, 51 residues in contact) was proposed to be the biological interface. Valdar and Thornton[19] reported that the biological contact of 1uby was not the largest contact but the second one. Alhough both of them are reasonable dimers (Kim Henrick, personal communication with), our result shows that 1uby (A:C, symmetry: -X, Y, -Z, 51 residues in contact) might be the more favored biological interface. The two possible biological dimer interfaces are shown in Figure 3. The real biological interface of 1uby needs to be validated by further experiments.

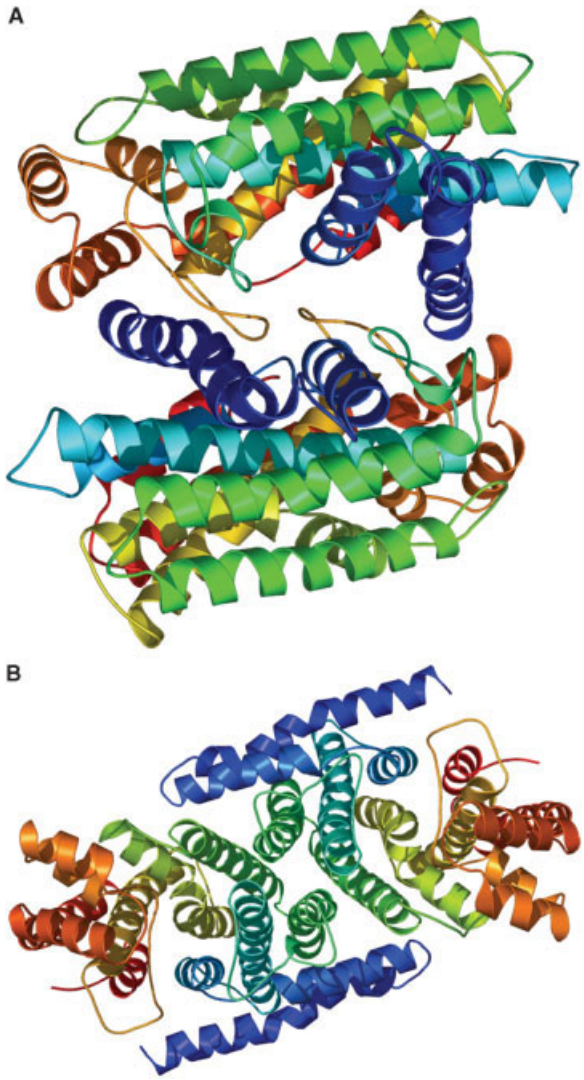

Fig. 3. Homodimeric crystal structure of farnesyl pyrophosphate synthetase (PDB code 1uby), two possible binding modes. **a:** 1uby A:B, symmetry: 1/2-X, Y, 3/4-Z, 49 residues in contact. **b:** 1uby A:C, symmetry: -X, Y, -Z, 51 residues in contact. Images were created by using PyMOL.[43] [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

PMFScore and Contact Size Is the Best Pairwise Combination for Distinguishing Biological Interfaces and Nonbiological Contacts

PMFScore and contact size are the best pairwise combination for distinguishing biological interfaces and nonbiological contacts. The results summarized in Table I show that PMFScore and contact size provide a good combination for discriminating biological from nonbiological contacts. As for predictor ① ③ listed in Table I, success rates are 79.1% of the monomers and 97.0% of the homodimers. In 2004, by analyzing a data set of 188 monomers and 122 homodimers, Bahadur et al.[22] reported that success rates were 77% of the monomers and 88% of the homodimers by a simple combination of the nonpolar interface area and the fraction of buried interface atoms. Although these success rates increase to 93–95% when the author applied a cutoff at RP = 1.5 to those misclassified cases at the first step. It is problematic when the author applied the cutoff to region M and D and treated them as the region U. Our predictor based on PMFScore and contact size is comparable with Bahadur's method based on fraction of buried atoms and nonpolar interface area.[22] Our success rates also increase to 92% for the homodimers and 81% for the monomers when the SCScore and packing density are taken into consideration.

The Combination of the Four Parameters Can Be Used for Distinguishing Biological and Nonbiological Contacts

The combination of the four parameters can be used to distinguish biological and nonbiological contacts. The only difference is SCScore between CTPScore (① ② ③ in Table I) and CFPScore (① ② ③ ④ in Table I). Ten cases (1lyn A:L, 1moq A:E, 1uby A:C, 2rsp A:G, 1bam A:D, 1gvp A:D, 1isa A:B, 1lyn A:B, 2ccy A:B, and 3ssi A:F) are misclassified by the CFPScore, and nine cases (1uby A:C, 1auo A:B, 1bam A:D, 1isa A:B, 1lyn A:B, 1uby A:B, 1xso A:B, 2ccy A:B, and 3ssi A:F) are misclassified by the CTPScore. Five interfaces (1lyn A:B, 1uby A:C, 1bam A:D, 2ccy A:B, and 3ssi A:F) are misclassified with both scores. Abalone sperm lysin (PDB 1lyn)[39] exists as a homodimer in solution, and the monomer is the active species. Shaw et al.[39] suggested that 1lyn (A:B, symmetry: X, Y, Z, 24 residues in contact) should be the biological interface. Our result shows that both 1lyn (A:L, symmetry: X, -Y, -Z, 21 residues in contact) and 1lyn (A:B, symmetry: X, Y, Z, 24 residues in contact) might be biological homodimers. The real biological interface of 1lyn needs to be validated by further experiments. Restriction endonuclease bamHI (PDB 1bam)[40] is a homodimer with a large cleft that could accommodate B-form DNA. The contact size of 1bam (A:D) is 18, which can be classified as biological interface using packing density and SCScore. Ferricytochrome C′ (PDB 2ccy)[41] from rhodospirillum molischianum belongs to electron transport (heme protein). It is apparent that complexes formed by electron transfer proteins have different properties than complexes formed by subunits of oligomeric protein. The interface

TABLE V. Performance of CFPScore and the best CTPScore on the protomers misclassified by Ponstingl et al.[27]

PDB code	Multimerric state (M/D)[a]	Correct (−/+)		
		PMFScore	the best CTPScore[b]	CFPScore[c]
1ako	M	−	−	−
1feh	M	−	−	−
1ckm	M	−	−	−
1avp	M	+	+	+
1ton	M	+	+	+
1jsg	D	+	+	+
1af5	D	−	+	+
1xso	D	−	−	+
1cp2	D	+	+	+
1auo	D	−	−	+
1slt	D	+	+	+

[a]M = monomer; D = dimer.
[b]The best CTPscore is ① ② ③ listed in Table II.
[c]− Represents misclassified; + represents correctly classified. A classification is correct only if all contacts in a crystal structure are correctly classified.

(3ssi A:F) of streptomyces subtilisin inhibitor has a low SCScore (92) and a high PMFScore (−30658).

The same data set was analyzed by Ponstingl et al.,[27] who tried to predict whether a given protomer was in a homodimer or monomer. Their pair potential based on atom-pair frequencies observed across hypothetical dimer interfaces in the 76 homodimers, classified wrongly 5 of 96 monomers and 7 of 76 homodimers. We applied the CFPScore to the 11 cases (1kbp was excluded in our data set) misclassified by Ponstingl's pair potential. Table V lists these 11 cases. CFPScore correctly classified all contacts, thereby also correctly predicting the multimeric state in eight of these protomers. The results show that the combination of the four parameters can be used for distinguishing biological and nonbiological contacts, especially in those cases misclassified by Ponstingl et al.[27] In the article by Nooren and Thornton,[42] PDB entry 1ckm was excluded on the basis of insufficient evidence for a biological monomer. In our calculation, PDB entry 1ako, 1feh, and 1ckm occur in the region of biological homodimers. Their oligomeric states need to be validated by further experiments.

As stated in Materials and Methods, unlike others,[21,27] we did not select the largest contacts, but we kept the crystal contacts with contact size > 10 for the study. This is because for some dimers, the largest interfaces are not the biological ones. To compare with the published results, we also did the analysis by using only the largest contacts. We found that 7 of the 74 homodimers and 16 of the 91 monomers are misclassified by CFPScore; 4 of the 74 homodimers and 18 of the 91 monomers are misclassified by SVM0; 9 of the 74 homodimers and 7 of the 91 monomers are misclassified by SVM1. Ponstingl et al.[27] used the difference in ΔASA between the largest and the second largest contact encountered in the crystals, and they found 8 dimers and 9 monomers are misclassified. The modified ASA score gave an accuracy of 88.9%. Valdar and Thornton[19] used the same data set by Ponstingl et al.[27] Because of the requirements of the method (sufficient sequence information available for residue conservation calculation), only 53 homodimers and 65 monomers are used. Alhough Valdar considered not only the largest contact, their data set was too small to be directly comparable with ours. Mintseris and Weng[21] achieved a success rate of 93% (misclassified 6 dimers as monomers and 6 monomers as dimers) for distinguishing between homodimers and crystal contacts by KDA with ACV. Their success rate was slightly higher than ours, but application on docking decoys was not reported.

CFPScore and SVM Model

Table I shows that the SVM models improve slightly over CFPScore to differentiate biological and nonbiological protein–protein interfaces. When CFPScore and SVM models are applied to the two docking decoys sets, CFPScore is significantly better than SVM models, which may come from the possible overlearning of SVM models toward crystal contacts. The CFPScore, a simple combination of four parameters, seems to better reflect the characteristics of protein–protein interfaces and can be applied to protein–protein docking and design studies.

CONCLUSION

To define a suitable scoring function that can be used in protein–protein interaction predictions, we have studied the performance of four parameters: contact size, packing density, geometric complementarity, and potential of mean force and their combinations to distinguish biological from nonbiological contacts for the protein homodimer and monomer data set. The results show that PMFScore is a better discriminator between biological and nonbiological interfaces comparable in size. When combining cutoff of PMFScore and one of the other three parameter's cutoffs, the error rates on monomers are decreased to 19.8–20.9%. The combination of PMFScore and contact size is the most powerful pairwise discriminator. Based on packing density, contact size, geometric complementarity, and potential of mean force, a combinatorial score (CFPScore) has

been developed, which gives the success rate of the homodimers discrimination of 96.6% and error rate of the monomers of 6.0% and 19.8% according to Valdar's and our definition, respectively. CFPScore was found to perform well with docking decoy sets. Compared with other statistical learning models, the cutoffs for the four parameters and their combinations are directly based on physical models, simple, and can be easily applied to protein–protein interface analysis and docking studies.

ACKNOWLEDGMENTS

The authors thank Dr. Hannes Ponstingl for kindly providing PITA program and data sets of homodimers and monomers, Lin Jiang for the atomic potential of mean force, Shi Tang for data preparation in the early stage of the project.

REFERENCES

1. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 2002;415:180–183.
2. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002;415:141–147.
3. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 2001;98:4569–4574.
4. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 2000;403:623–627.
5. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature 1999;402:86–90.
6. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. Science 1999;285:751–753.
7. Goh CS, Cohen FE. Co-evolutionary analysis reveals insights into protein-protein interactions. J Mol Biol 2002;324:177–192.
8. Aloy P, Bottcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, Gavin AC, Bork P, Superti-Furga G, Serrano L, Russell RB. Structure-based assembly of protein complexes in yeast. Science 2004;303:2026–2029.
9. Deng M, Sun F, Chen T. Assessment of the reliability of protein-protein interactions and protein function prediction. Pac Symp Biocomput 2003:140–151.
10. Sprinzak E, Sattath S, Margalit H. How reliable are experimental protein-protein interaction data? J Mol Biol 2003;327:919–923.
11. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
12. Liang S, Liu Z, Li W, Ni L, Lai L. Construction of protein binding sites in scaffold structures. Biopolymers 2000;54:515–523.
13. Smith GR, Sternberg MJ. Prediction of protein-protein interactions by docking methods. Curr Opin Struct Biol 2002;12:28–35.
14. Janin J. Assessing predictions of protein-protein interaction: the CAPRI experiment. Protein Sci 2005;14:278–283.
15. Chothia C, Janin J. Principles of protein-protein recognition. Nature 1975;256:705–708.
16. Janin J, Chothia C. The structure of protein-protein recognition sites. J Biol Chem 1990;265:16027–16030.
17. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
18. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177–2198.
19. Valdar WS, Thornton JM. Conservation helps to identify biologically relevant crystal contacts. J Mol Biol 2001;313:399–416.
20. Valdar WS, Thornton JM. Protein–protein interfaces: analysis of amino acid conservation in homodimers. Proteins 2001;42:108–124.
21. Mintseris J, Weng Z. Atomic contact vectors in protein–protein recognition. Proteins 2003;53:629–639.
22. Bahadur RP, Chakrabarti P, Rodier F, Janin J. A dissection of specific and non-specific protein-protein interfaces. J Mol Biol 2004;336:943–955.
23. Robert CH, Janin J. A soft, mean-field potential derived from crystal contacts for predicting protein-protein interactions. J Mol Biol 1998;283:1037–1047.
24. Jiang L, Lai L. CH…O hydrogen bonds at protein-protein interfaces. J Biol Chem 2002;277:37732–37740.
25. Jiang L, Gao Y, Mao F, Liu Z, Lai L. Potential of mean force for protein–protein interaction studies. Proteins 2002;46:190–196.
26. Jackson RM, Gabb HA, Sternberg MJ. Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. J Mol Biol 1998;276:265–285.
27. Ponstingl H, Henrick K, Thornton JM. Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 2000;41:47–57.
28. Ponstingl H KT, Thornton JM. Automatic inference of protein quaternary structure from crystals. J Appl Crystallogr 2003;36:1116–1122.
29. Hubbard STJ. Naccess. London: Department of Biochemistry Molecular Biology, University College, 1993.
30. Richards FM. The interpretation of protein structures: total volume, group volume distributions and packing density. J Mol Biol 1974;82:1–14.
31. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci USA 1992;89:2195–2199.
32. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
33. Lu H, Lu L, Skolnick J. Development of unified statistical potentials describing protein-protein interactions. Biophys J 2003;84:1895–1901.
34. Zhang C, Liu S, Zhou H, Zhou Y. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. Protein Sci 2004;13:400–411.
35. Li L, Chen R, Weng Z. RDOCK: refinement of rigid-body protein docking predictions. Proteins 2003;53:693–707.
36. Reed BC, Rilling HC. Substrate binding of avian liver prenyltransferase. Biochemistry 1976;15:3739–3745.
37. Tarshis LC, Proteau PJ, Kellogg BA, Sacchettini JC, Poulter CD. Regulation of product chain length by isoprenyl diphosphate synthases. Proc Natl Acad Sci USA 1996;93:15018–15023.
38. Henrick K, Thornton JM. PQS: a protein quaternary structure file server. Trends Biochem Sci 1998;23:358–361.
39. Shaw A, Fortes PA, Stout CD, Vacquier VD. Crystal structure and subunit dynamics of the abalone sperm lysin dimer: egg envelopes dissociate dimers, the monomer is the active species. J Cell Biol 1995;130:1117–1125.
40. Newman M, Strzelecka T, Dorner LF, Schildkraut I, Aggarwal AK. Structure of restriction endonuclease bamhi phased at 1.95 A resolution by MAD analysis. Structure 1994;2:439–452.
41. Finzel BC, Weber PC, Hardman KD, Salemme FR. Structure of ferricytochrome c′ from Rhodospirillum molischianum at 1.67 A resolution. J Mol Biol 1985;186:627–643.
42. Nooren IM, Thornton JM. Structural characterisation and functional significance of transient protein-protein interactions. J Mol Biol 2003;325:991–1018.
43. DeLano WL. The case for open-source software in drug discovery. Drug Discov Today 2005;10:213–217.