ELSEVIER

Biophysical Chemistry 129 (2007) 1 – 10

Biophysical Chemistry

http://www.elsevier.com/locate/biophyschem

Complex-type-dependent scoring functions in protein–protein docking

Chun Hua Li [a,1], Xiao Hui Ma [a,b,1], Long Zhu Shen [a], Shan Chang [a], Wei Zu Chen [a], Cun Xin Wang [a,*]

[a] College of Life Science and Bioengineering, Beijing University of Technology, Beijing 100022, People's Republic of China
[b] Department of Pharmaceutical and Biomedical Sciences, University of George, Athens, GA 30602, USA

Received 14 February 2007; received in revised form 24 April 2007; accepted 25 April 2007
Available online 8 May 2007

Abstract

A major challenge in the field of protein–protein docking is to discriminate between the many wrong and few near-native conformations, i.e. scoring. Here, we introduce combinatorial complex-type-dependent scoring functions for different types of protein–protein complexes, protease/inhibitor, antibody/antigen, enzyme/inhibitor and others. The scoring functions incorporate both physical and knowledge-based potentials, i.e. atomic contact energy (ACE), the residue pair potential (RP), electrostatic and van der Waals' interactions. For different type complexes, the weights of the scoring functions were optimized by the multiple linear regression method, in which only top 300 structures with ligand root mean square deviation (L_RMSD) less than 20 Å from the bound (co-crystallized) docking of 57 complexes were used to construct a training set. We employed the bound docking studies to examine the quality of the scoring function, and also extend to the unbound (separately crystallized) docking studies and extra 8 protein–protein complexes. In bound docking of the 57 cases, the first hits of protease/inhibitor cases are all ranked in the top 5. For the cases of antibody/antigen, enzyme/inhibitor and others, there are 17/19, 5/6 and 13/15 cases with the first hits ranked in the top 10, respectively. In unbound docking studies, the first hits of 9/17 protease/inhibitor, 6/19 antibody/antigen, 1/6 enzyme/inhibitor and 6/15 others' complexes are ranked in the top 10. Additionally, for the extra 8 cases, the first hits of the two protease/inhibitor cases are ranked in the top for the bound and unbound test. For the two enzyme/inhibitor cases, the first hits are ranked 1st for bound test, and the 119th and 17th for the unbound test. For the others, the ranks of the first hits are the 1st for the bound test and the 12th for the 1WQ1 unbound test. To some extent, the results validated our divide-and-conquer strategy in the docking study, which might hopefully shed light on the prediction of protein–protein interactions.

Keywords: Binding affinity; Scoring function; Protein–protein docking

1. Introduction

Protein–protein interaction is the basis of many biological regulations. Knowledge of 3-dimensional (3D) protein–protein structures is important for an adequate description of protein–protein interactions. However, large macromolecular assemblies are a major challenge for structural biology. The amount of experimental structures of protein–protein complexes is relatively quite small and the cost is very expensive. Thus, a combination of protein modeling and experimental structure determination increases knowledge of structure-based analysis of the protein–protein interaction network [1–4]. As a part of molecular modeling, docking algorithms are designed to model protein–protein complexes based on the component structures.

Docking algorithms have progressed in recent years, which can dock unbound (separately crystallized) proteins to obtain the structure of the complex with small structural changes accompanying complexation [5–19]. The accuracy and reliability of docking algorithms still need to be assessed if they are to become widely used. This depends on docking algorithms with an efficient procedure to generate potential structures and a good scoring function to distinguish the near-native structures from a large number of non-native ones. The known scoring functions include surface complementarity (SC) [5,6], surface complementarity together with an electrostatic filter [20,21], knowledge-based statistical potential such as atomic contact energy (ACE) [22], the residue pair potential (RP) [23] and DFIRE [24]. Some combinatorial functions are used in docking

* Corresponding author.
E-mail address: cxwang@bjut.edu.cn (C.X. Wang).
[1] Both are first authors.

0301-4622/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2007.04.014

prediction [17,25]. Although the present scoring functions have achieved some success, none has been proved to be robust enough to all types of protein–protein complexes.

Some investigations have revealed that enzyme/inhibitor, antibody/antigen and other protein complexes present some differences in the residue composition, hydrophobicity and electrostatics at the interface [26–29]. Our previous work showed that the complex-type-dependent filtering strategies can improve the docking predictions [30]. Here we introduced complex-type-dependent combinatorial scoring functions for protease/inhibitor, enzyme/inhibitor, antibody/antigen and other complexes. The scoring functions incorporate both physical and knowledge-based potentials, i.e. ACE [22], RP [23], electrostatic and van der Waals' interactions. For different type complexes, the weights of the scoring functions were optimized by the multiple linear regression method, in which only top 300 structures with ligand root mean square deviation (L_RMSD) less than 20 Å from the bound (co-crystallized) docking were used to construct the training set. We employ the bound docking studies to examine the quality of the scoring function, and also extend to the unbound docking studies. Finally, the complex-type-dependent combinatorial scoring functions are tested with extra 8 protein–protein complexes.

2. Materials and methods

2.1. Dockings sets and decoys generation

59 Targets are selected from the Benchmark 1.0 [31], extra 4 cases (1E6J, 1BVN, 1AY7, 1VFB) are from Benchmark 2.0 [32] and two complexes 1CA0 and 1TAW are from the literature [33]. Both bound and unbound docking processes were performed for all complexes by the FTDock program [5]. For each case, 30,000 structures respectively from bound and unbound docking were generated with the shape complementarity criterion in the searching stage. The top 300 bound docked decoys with the L_RMSD less than 20 Å were utilized to fit the weights of the scoring functions. L_RMSD values were computed over backbone atoms (N, C, CA, O) of the ligands after the receptors of the decoy and the native structure were superimposed. All bound and unbound docked decoys were used to test the discriminative abilities of the scoring functions.

2.2. Biological information filter

For antibody/antigen, complementarity determining regions (CDRs) usually bind the antigen epitope. Therefore, the CDRs

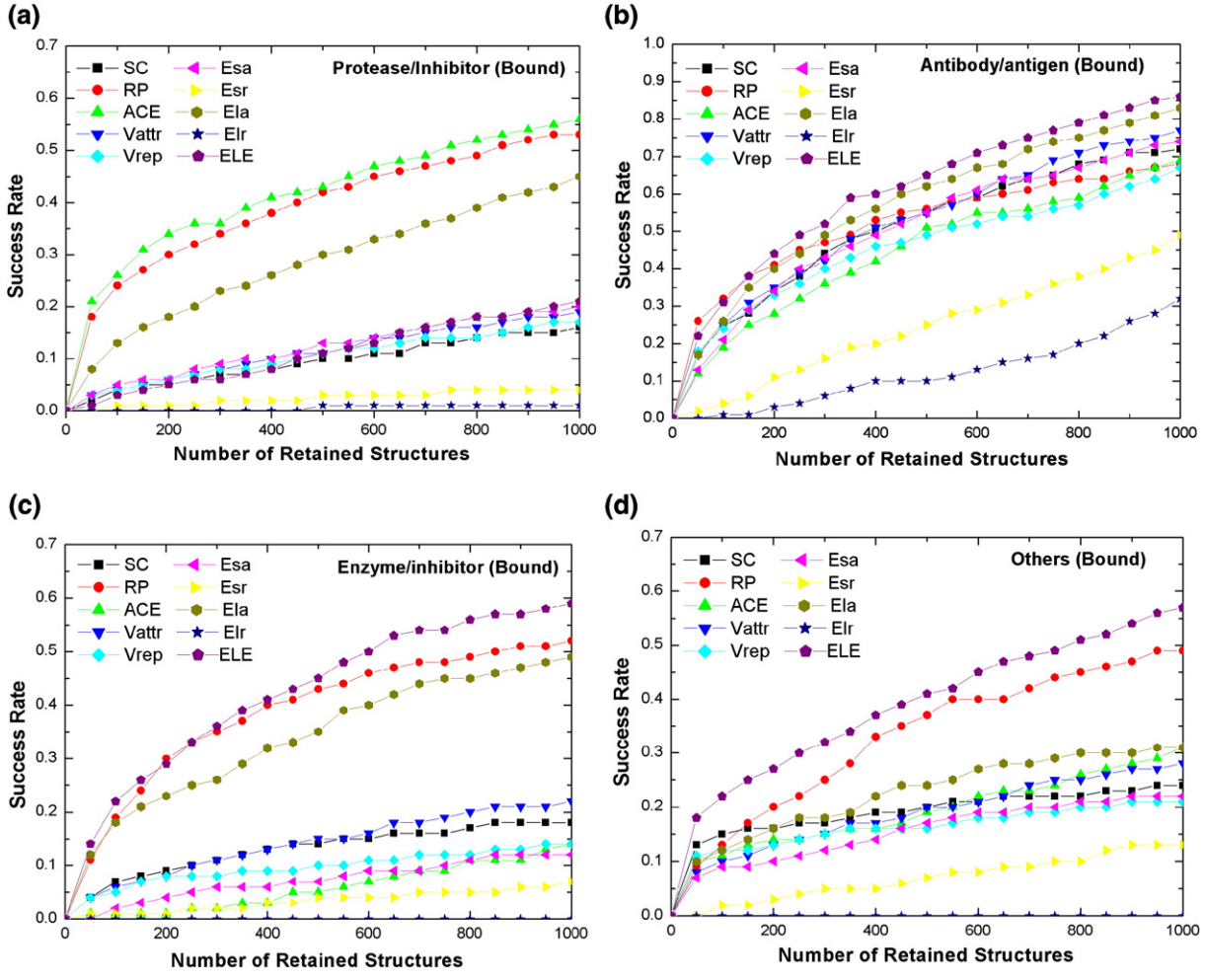

Fig. 1. Success rate comparisons among scoring components for bound dockings of the four types of complexes. (a) Protease/inhibitor; (b) antigen/antibody; (c) enzyme/inhibitor; (d) others.

Table 1
The six fitted scoring functions with different combinations of scoring components for four types of complexes

Scores[a]	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
Protease/inhibitor								
Score1	−0.267	0.107	0	0	0	0	0	0
Score2	−0.310	0	0.277	0	0	0	0	0
Score3	−0.299	0	0	0.0094	0	0	0	0
Score4	−0.308	0	0.338	0.131	0	0	0	0
Score5	−0.286	0.209	−0.577	0.126	0.155	0.126	0.160	0.157
Score6	0	0.115	0.0762	0.121	0.0808	0.0842	0.0982	0.0962
Antigen/antibody								
Score1	−0.254	0.0228	0	0	0	0	0	0
Score2	−0.227	0	0.416	0	0	0	0	0
Score3	−0.232	0	0	0.0429	0	0	0	0
Score4	−0.223	0	0.485	0.154	0	0	0	0
Score5	−0.364	0.082	0.248	0.281	0.137	0.125	0.139	0.137
Score6	0	0.0275	0.303	0.142	0.0585	0.0751	0.0766	0.0730
Enzyme/inhibitor								
Score1	−0.44212	0.0697	0	0	0	0	0	0
Score2	−0.473	0	0.350	0	0	0	0	0
Score3	−0.452	0	0	0.0498	0	0	0	0
Score4	−0.471	0	0.392	0.101	0	0	0	0
Score5	−0.670	0.147	0.084	0.086	0.043	0.044	−0.018	−0.021
Score6	0	0.079454	0.2165	0.0856	0.0601	0.0772	0.1235	0.120
Others								
Score1	−0.320	0.0756	0	0	0	0	0	0
Score2	−0.379	0	0.370	0	0	0	0	0
Score3	−0.360	0	0	0.0629	0	0	0	0
Score4	−0.376	0	0.416	0.0929	0	0	0	0
Score5	−0.156	0.123	0.382	0.048	0.094	0.104	0.130	0.128
Score6	0	0.0887	0.288	0.0989	0.0413	0.05807	0.0817	0.0791

[a] w_1–w_8 are the weights of energy items in Eq. (1).

were translated into distance constraints to force these particular residues of antibody to be close to antigen. The docked modes with any atom of CDRs within 6 Å of any atom of the antigen were retained. For protease/inhibitor, enzyme/inhibitor and others, the biological interface is much dependent on the complex itself. So, not any filter was performed for these three types of complexes.

2.3. Scoring functions

Scoring functions are summed linearly, including the residue pair potential (E_{RP}), desolvation (E_{ACE}), van der Waals' attractive ($E_{\mathrm{vdw}}^{\mathrm{attr}}$) and repulsive ($E_{\mathrm{vdw}}^{\mathrm{rep}}$) energies, electrostatic short-range attractive ($E_{\mathrm{ele}}^{\mathrm{sa}}$) and repulsive ($E_{\mathrm{ele}}^{\mathrm{sr}}$), electrostatic long-range attractive ($E_{\mathrm{ele}}^{\mathrm{la}}$) and repulsive ($E_{\mathrm{ele}}^{\mathrm{lr}}$) energies. These scoring items are chosen as they together provide a relatively comprehensive representation of the energies in protein complex formation. The general combinatorial scoring function is formulated below:

$$\begin{aligned}\mathrm{Score} = {} & w_1E_{\mathrm{RP}} + w_2E_{\mathrm{ACE}} + w_3E_{\mathrm{vdw}}^{\mathrm{attr}} + w_4E_{\mathrm{vdw}}^{\mathrm{rep}} + w_5E_{\mathrm{ele}}^{\mathrm{sa}} \\ & + w_6E_{\mathrm{ele}}^{\mathrm{sr}} + w_7E_{\mathrm{ele}}^{\mathrm{la}} + w_8E_{\mathrm{ele}}^{\mathrm{lr}} ,\end{aligned} \tag{1}$$

where w_1–w_8 are the weights of energy items, which are obtained using multiple linear regression method by fitting between the scoring items and L_RMSD values.

The desolvation term is based on the ACE model [22]

$$E_{\mathrm{ACE}} = \sum_i \sum_j e_{ij} , \tag{2}$$

where e_{ij} denotes the ACE between atoms i and j, and the sum is taken over all atom pairs less than 6 Å apart. The electrostatic energy calculation adopts the Coulomb model used in RosettaDock [15]

$$E_{\mathrm{ele}} = \frac{332q_iq_j}{\varepsilon_r \hat{r}_{ij}} = \frac{332q_iq_j}{\hat{r}_{ij}^2}, \tag{3}$$

where $\hat{r}$ = max (r_{ij}, 3 Å) to avoid singularities as $r_{ij} \to 0$. Interactions are separated into attractive and repulsive categories as well as short-range ($r_{ij} < 5$ Å) and long-range ($r_{ij} \geq 5$ Å) categories. Van der Waals' attractive and repulsive items are calculated with the modified Lennard–Jones 6–12 potential [34]

$$V_{\mathrm{attr}} = \sum_i \sum_j \varepsilon_{ij} \left[\left(\frac{r_{\mathrm{m}\ ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{\mathrm{m}\ ij}}{r_{ij}} \right)^{6} \right] \quad \text{if} \quad r_{ij} \geq 0.89 r_{m,ij} \tag{4}$$

$$V_{\mathrm{rep}} = \sum_i \sum_j 10.0 \times \left(1 - \frac{r_{ij}}{0.89 \times r_{m,ij}} \right) \text{ if } r_{ij} < 0.89 r_{m,ij} \tag{5}$$

Where $r_{m,ij}$ is the sum of Lennard–Jones radii of atoms i and j of the two interactive molecules. r_{ij} is the distance between the two atoms.

ε_{ij} is the square root of the product of the well depths. Van der Waals' item is continuous at $r_{ij}=0.89\ r_{m,ij}$. The values of all parameters are taken from the CHARMM19 parameter set [35].

2.4. Assessment criteria of scoring functions

Here, "success rate" was used to investigate the contribution of each scoring component to the filtering result. It is defined as the average ratio of the number of the hit structures retained by a filter to that of all hit structures in a certain number of decoys in all cases. In addition, the rank of the first hit structure and the number of hits in the top ten docked modes of scoring list were used to evaluate the discriminative abilities of the complex-type-dependent scoring functions. The hit structure is defined as the one with L_RMSD < 10 Å [36].

3. Results and discussion

3.1. Protein complexes classification

Early studies have pointed out that the principles governing the interactions involved in protein–protein recognition present significant differences for the different types of complexes. The physico-chemical characteristics of the binding interface vary with different types of complexes [27]. The primary principle of our classification is based on the biological functions, i.e. enzyme/inhibitor, antibody/antigen and others. Furthermore we separated protease/inhibitors from enzyme/inhibitors because these two complexes hold distinctive features at the interfaces. There exist more hydrophobic interactions at the interface for the former than the latter. Protease/inhibitor complexes have more non-polar interface rate (55.5% in average), whereas the enzyme/inhibitor complexes have more polar interface rate (57.2% in average). The solvent accessibility area has been calculated with the program NACCESS [37].

3.2. Optimizing combinatorial scoring functions

To investigate the contribution of each scoring component to the filtering result, 10 energy filters were applied to the bound docked decoys (shown in Fig. 1). For protease/inhibitor, ACE and RP filters can retain more hits than the other filters, and furthermore, the ACE filter performs slightly better than the RP filter. For antibody/antigen, enzyme/inhibitor and others, ELE

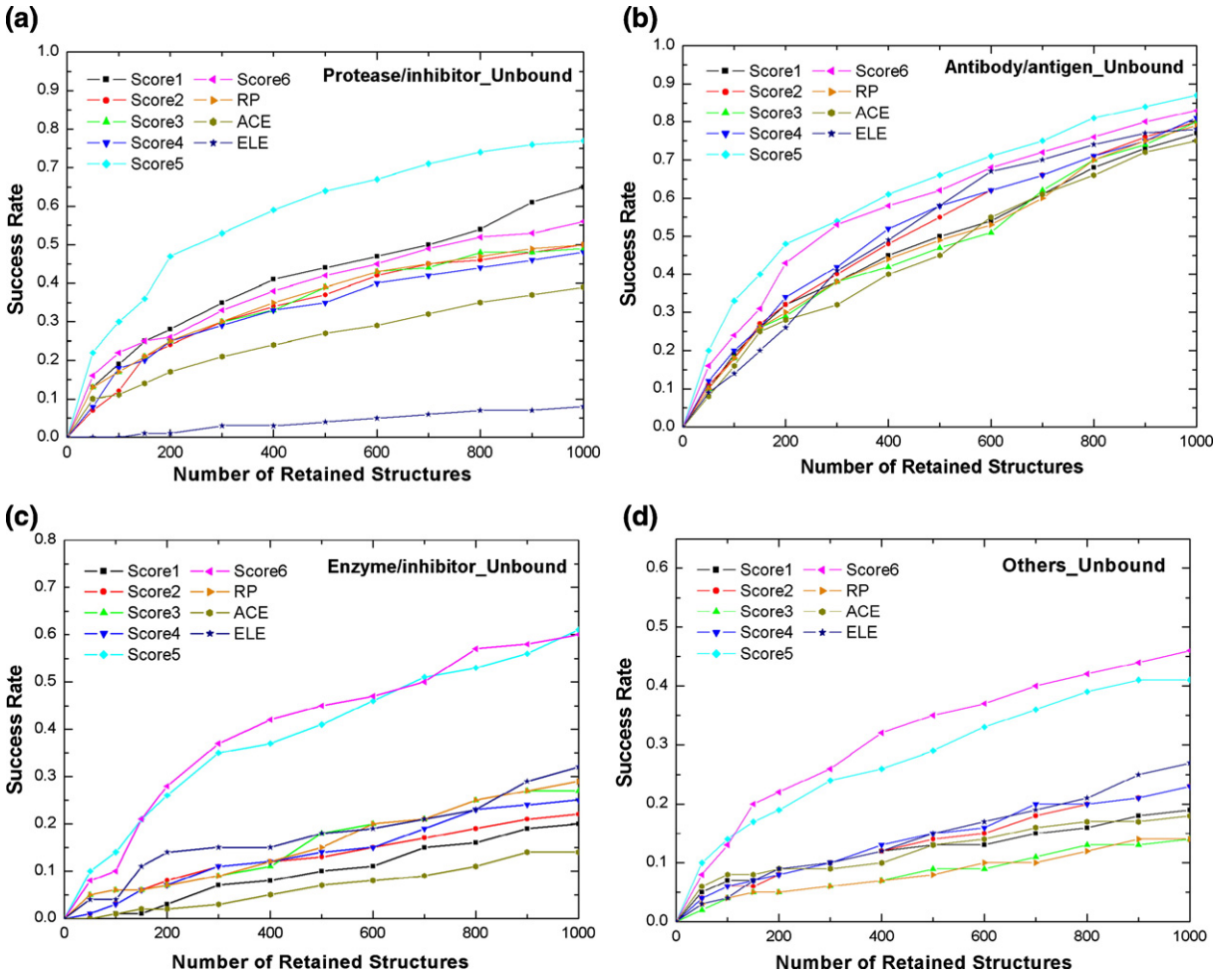

Fig. 2. Success rate comparisons among the six fitted scoring functions with different combinations of scoring components and three single components RP, ACE and ELE for unbound dockings of the four types of complexes. (a) Protease/inhibitor; (b) antibody/antigen; (c) enzyme/inhibitor; (d) others.

Table 2
The discriminative capacities of four complex-type-dependent scoring functions

Training set	Bound			Unbound		
	L_RMSD of 1st hit [a]	Rank of 1st hit [b]	Number of hits [c]	L_RMSD of 1st hit [a]	Rank of 1st hit [b]	Number of hits [c]
Protease/inhibitor (17)						
1ACB	4.27	1	9	4.08	1	9
1AVW	1.76	1	10	9.23	567	0
1BRC	4.78	1	3	7.12	1	3
1BTH [d]	2.22	1	10	8.45	151	0
1CGI	1.37	1	10	7.38	47	0
1CHO	3.89	1	9	3.04	1	8
1CSE	1.40	4	5	2.48	8	2
1PPE	1.64	1	10	1.05	1	10
1STF	1.97	1	7	1.48	1	8
1TAB	1.82	3	2	1.21	14	0
1TGS	1.72	1	10	8.81	69	0
2KAI	4.37	1	8	3.26	46	0
2PTC	1.54	1	10	8.38	4	4
2SIC	6.67	1	10	5.22	1	3
2SNI	4.12	1	10	4.04	16	0
2TEC	4.05	4	7	3.15	1	6
4HTC	1.33	1	5	7.47	814	0
Antibody/antigen (19)						
1AHW	3.25	3 (1)	3 (6)	6.34	373 (1)	0 (0)
1BQL	2.05	1 (1)	3 (3)	5.32	191 (92)	0 (0)
1BVK	7.92	10 (4)	1 (1)	9.83	26 (10)	0 (1)
1DQJ	1.36	3 (2)	3 (6)	8.55	1418 (131)	0 (0)
1EO8	4.63	1 (1)	5 (7)	9.36	328 (63)	0 (0)
1FBI	2.04	1 (1)	7 (7)	6.79	943 (51)	0 (0)
1IAI	1.84	2 (2)	5 (7)	4.96	18 (13)	0 (0)
1JHL	1.58	20 (17)	0 (0)	2.10	34 (26)	0 (0)
1KXQ	1.59	1 (1)	4 (7)	1.42	2 (1)	2 (3)
1KXT	1.59	8 (2)	1 (3)	1.25	296 (10)	0 (1)
1KXV	2.05	2 (1)	3 (5)	4.47	109 (2)	0 (1)
1MEL	2.97	2 (1)	3 (8)	3.36	405 (81)	0 (0)
1MLC	1.91	2 (1)	3 (4)	3.57	558 (109)	0 (0)
1NCA	1.59	2 (2)	2 (2)	3.68	530 (49)	0 (0)
1NMB	2.75	15 (1)	0 (1)	5.17	563 (22)	0 (0)
1QFU	2.01	1 (1)	9 (10)	3.97	6 (2)	1 (4)
1WEJ	4.85	129 (1)	0 (5)	5.85	55 (50)	0 (6)
2JEL	1.82	1 (1)	3 (5)	5.60	85 (50)	0 (0)
2VIR	1.46	253 (91)	0 (0)	1.81	43 (23)	0 (0)
Enzyme/inhibitor (6)						
1BRS	3.06	1	10	6.19	18	0
1DFJ	1.46	1	9	9.02	3	1
1FSS	1.63	1	7	9.21	474	0
1MAH	3.32	2	2	6.33	113	0
1UDI	1.91	52	0	9.76	384	0
1UGH	1.65	2	4	9.87	510	0
Others (15)						
1A0O	4.27	1	5	7.02	9	1
1ATN	4.06	1	8	2.00	1	4
1AVZ	4.07	3	1	–	–	–
1EFU [d]	1.62	1	10	9.41	2	1
1FIN [d]	1.41	1	10	–	–	–
1FQ1 [d]	1.78	7	2	4.22	1552	0
1GLA	1.73	3	2	–	–	–
1IGC	9.76	524	0	9.92	323	0
1KKL [d]	1.97	5	2	9.55	69	0
1L0Y	7.39	8590	0	4.53	19177	0
1SPB	3.79	1	3	6.65	1	4
2BTF	1.59	1	8	8.96	1	4
2MTA	7.5	1	4	8.47	271	0
2PCC	1.55	1	6	9.80	5	2
3HHR [d]	7.75	3	4	8.05	2700	0
Test set	Bound			Unbound		
	L_RMSD of 1st hit [a]	Rank of 1st hit [b]	Number of hits [c]	L_RMSD of 1st hit [a]	Rank of 1st hit [b]	Number of hits [c]
Protease/inhibitor (2)						
1CA0	3.15	1	10	5.90	1	7
1TAW	1.23	1	10	3.19	1	6
Antibody/antigen (2)						
1VFB	2.08	15 (10)	0 (1)	9.28	80 (39)	0 (0)
1E6J	1.88	21 (17)	0 (0)	8.62	111 (19)	0 (0)
Enzyme/inhibitor (2)						
1AY7	1.68	1	5	7.39	119	0
1BVN	3.98	1	9	9.83	17	0
Others (2)						
1GOT [d]	1.51	1	10	–	–	–
1WQ1	9.94	1	3	9.11	12	0

– No structures with L_RMSD < 10 Å obtained in 30,000 docked decoys.
() Scoring results after filtering. Antibody/antigen cases were filtered with CDRs information.
[a] L_RMSD of the first hit.
[b] The rank of the first hit.
[c] The number of hits within top 10 scores.
[d] Difficult cases referred in Benchmark 1.0.

filter shows better performance. To be notable, the electrostatic attractive and repulsive components appear different filtering potentials for four types of complexes.

To train a good scoring function to distinguish the near-native structures from a large number of non-native ones, different combinations of scoring components were tried and the weights of the scoring functions were fitted with the top 300 bound docked decoys with the L_RMSD < 20 Å. The decoys with L_RMSD < 10 Å appear the near-correct interfaces of each partner [36] and were taken as positive samples. On the other hand, the decoys with L_RMSD in 10–20 Å were used as negative samples to fit scoring function weights. For the docked decoys with the L_RMSD > 20 Å, the interfaces are more incorrect, thus inducing bias in deriving the weights of the short distance terms. Therefore, they were not considered in the linear regression. Table 1 shows the six fitted scoring functions with different combinations of scoring components for every type of complexes. Fig. 2 gives the change of success rate with the number of retained structures for the six fitted scoring functions, RP, ACE and ELE in unbound dockings of the four types of complexes. From Fig. 2, Score5, the combination of all scoring components, appears higher prediction success rate than the other scores for protease/inhibitor and antibody/antigen. For enzyme/inhibitor and others, Score5 and Score6 (the combination of all scoring components except for RP) display similar average prediction ability. Therefore, Score5 was used as the complex-type-dependent scoring function. Additionally, compared with the scoring function RP used in FTDock, Score5 shows an obvious advantage at discriminating the hit structure.

3.3. Combinatorial scoring functions performance

Table 2 reports the scoring results of the complex-type-dependent scoring functions for all four types of complexes. For the bound docking results, there are 51 of 57 cases (16/17

protease/inhibitor, 18/19 antibody/antigen, 6/6 enzyme/inhibitor, 11/15 others) with L_RMSD of the first hits less than 5 Å, and 51 cases (17 protease/inhibitor, 17 antibody/antigen, 5 enzyme/inhibitor, 12 others) with the first hit ranked above fifth. There are 100% cases with the first hit ranked in the top 10 for protease/inhibitor, 79% cases for antibody/antigen, 83% for enzyme/inhibitor and 87% for others. The analysis for antibody/antigen complexes is from the results without CDRs filtering. After the decoys are filtered with the CDRs information, the scoring results are improved apparently. There are 89% cases with the first hit ranked in the top 10. The first hits of 12 cases are ranked first, which increases by 6 compared with the rank

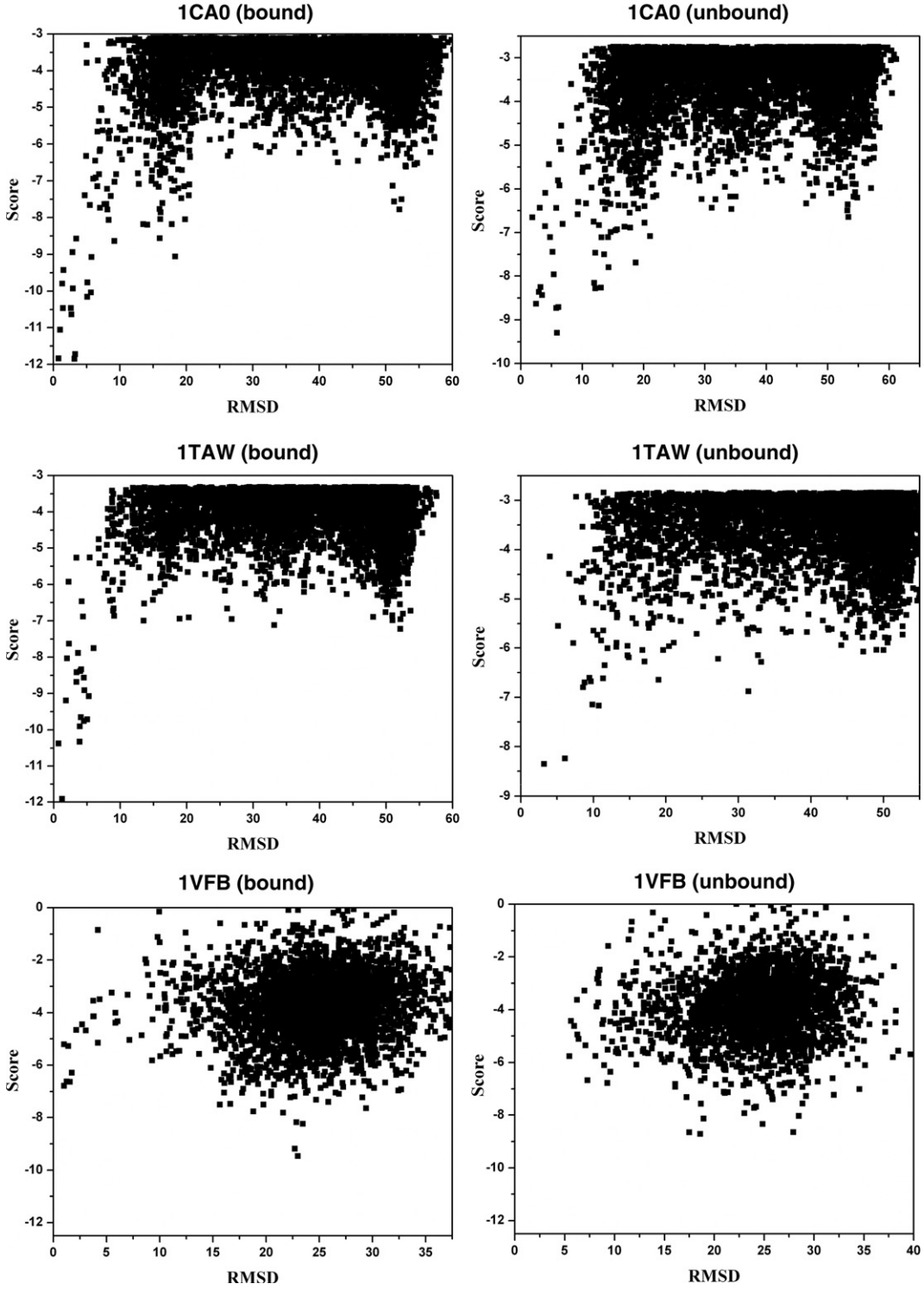

Fig. 3. Plots show L_RMSD versus scores for top 5000 decoys. Bound indicates that decoys are taken from the docking starting from two co-crystallized structures; unbound indicates that decoys are from the docking starting from one or two separately crystallized structures.

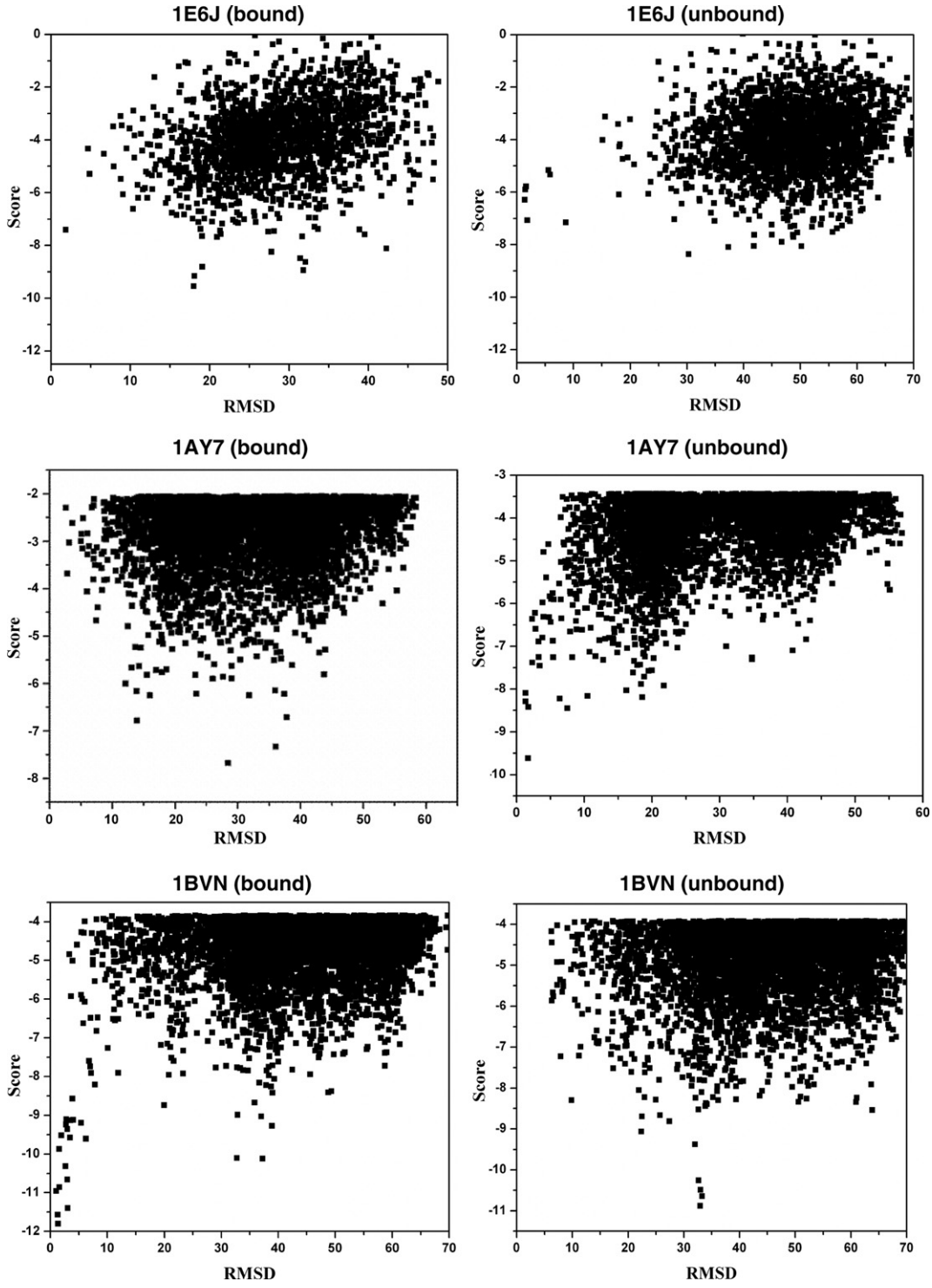

Fig. 3 (*continued*).

without filtering. Additionally, the number of hits in the top ten scores is raised for the most cases. This verifies the importance of biological information in the docking prediction, which agrees well with the previous investigations [19].

For unbound docking studies, L_RMSD of the first hits are bigger than that of the bound docked structures for most cases. There are 9 protease/inhibitor, 10 antibody/antigen, 0 enzyme/inhibitor and 3 others with L_RMSD of the first hits less than 5 Å, and 22 (9 protease/inhibitor, 6 antibody/antigen, 1 enzyme/inhibitor, 6 others) of 57 cases with near-native conformations found within top 10 scores. For the cases of others' type 1AVZ, 1FIN and 1GLA, no docked structures with L_RMSD less than 10 Å were obtained at sampling stage. For most of the four types of complexes, the rank of the 1st hit and the number of hits in

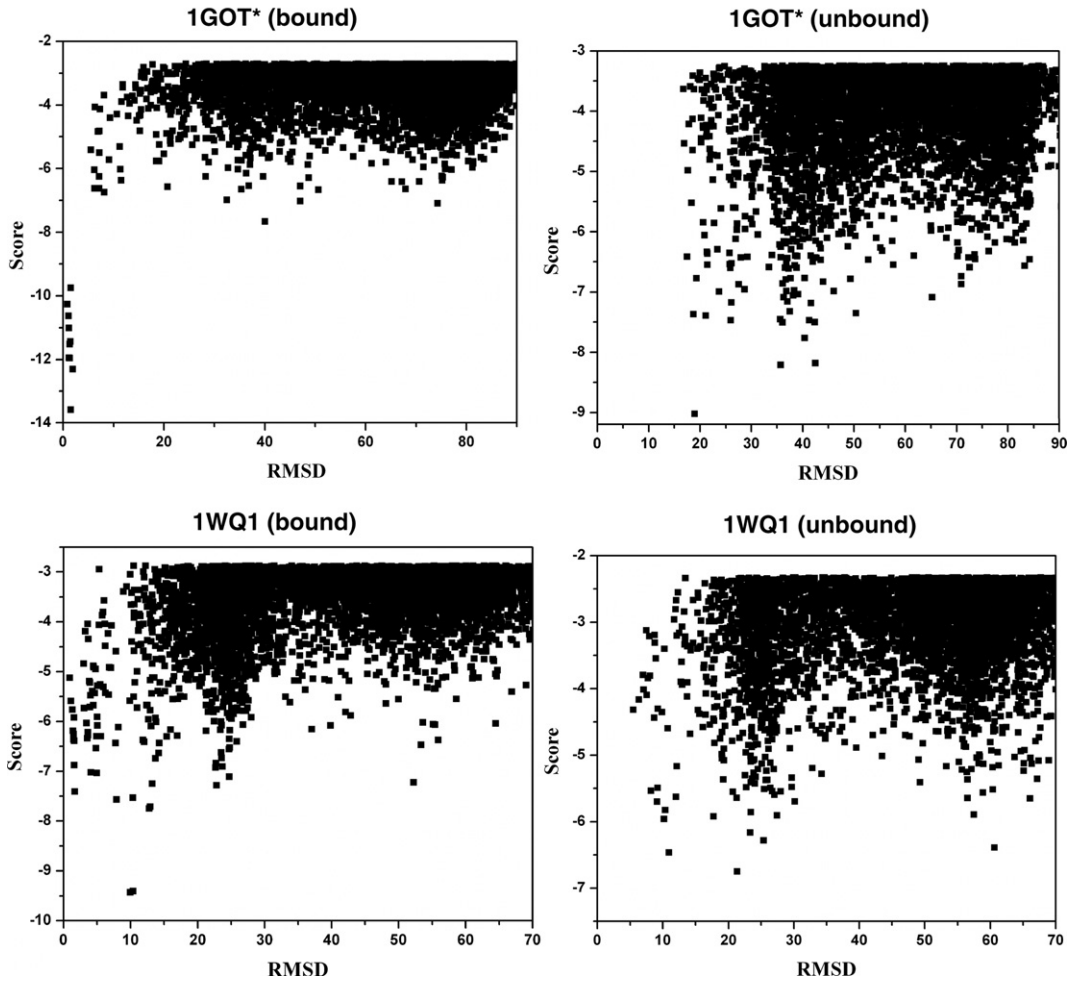

Fig. 3 (*continued*).

top 10 scores in unbound docking are not as good as that in bound docking. Evidently, one reason is that in contrast to bound docking, there are fewer hits obtained at sampling stage for unbound docking, which is not favorable for the performance of scoring functions. Currently, in protein–protein docking, there are still no good methods to consider the molecular conformational large change upon complex formation, which results into some failures in unbound docking. Just due to this reason, there are seven cases [16] referred as difficult in Benchmark 1.0, in which six (1EFU, 1FIN, 1FQ1, 1KKL, 3HHR and 1GOT) fall into the others' type and one (1BTH) belongs to the protease/inhibitor. From Table 2, there are six different cases (1BTH, 1EFU, 1FIN, 1FQ1, 1KKL and 3HHR) whose first hit structures in bound test are all ranked in the top 10. However, ranks of the 1st hit for unbound docked decoys are 151, 2, 1552, 69, 2700 for 1BTH, 1EFU, 1FQ1, 1KKL and 3HHR. For 1FIN, no hit structures are obtained at sampling stage. Analyzing the reason for not good result in the unbound score test, we think that too many overlaps at the interface of the docked decoy result in error of scoring. Structural optimization of the docked decoy will help improve the scoring effect. This work is currently underway.

Additionally, the extra 8 cases have been chosen to test the scoring functions. For the two protease/inhibitor cases, the first hits are ranked the top for bound test and unbound test. For antibody/antigen, the better scoring results have been observed after the filtering of decoys is performed based on CDRs. The unbound results are also acceptable with the first hit ranked the 39th and the 19th. For the two enzyme/inhibitor cases, the first hits are ranked 1st for bound test, and the 119th and the 17th for unbound test. For the two others' cases, the ranks of the first hits are the 1st for bound test and the 12th for 1WQ1 unbound test. Fig. 3 shows eight cases results from bound and unbound studies. For each target, the score values of the top 5000 decoys are plotted as a function of the L_RMSD values. If the scoring functions are successful, the lowest scoring structures will have the smallest L_RMSD to appear a score "funnel." For all protein–protein types, score funnels are apparent in the bound docking study, and the widths of the funnels range from 0 to 20 Å. Score funnels are also observed in the unbound docking of protease/inhibitor. Although score funnels are not clearly observed in the unbound docking of the other three complexes types, 1st hit from the bound docking are ranked first in the unbound

docking decoys. It is clear that generating near-correct conformations in the searching stage is a major challenge for the unbound docking.

4. Conclusion

Developing robust scoring functions is one of the main issues in complex structure prediction. Our goal is to optimize effective scoring functions which can distinguish the near-native structures from non-native ones for different types of protein–protein complexes. Through dividing complexes into four categories and designing the scoring function with the regression method for each type, we have finally got four combinatorial functions which exhibit certain aptitude to select hit structures from all docked solutions. For the four classes, our scoring functions show relatively good abilities in distinguishing hit structures. This result validates our divide-and-conquer strategy in the study of designing scoring functions. Additionally, with the help of more sufficient consideration of protein flexibility, the better effect of scoring is expected to be advanced. We hope this trial might provide some insight into the future scoring function development.

Acknowledgments

This work was supported by the Chinese Natural Science Foundation (grant nos. 30400087, 10574009 and 30170230), the Specialized Research Fund for the Doctoral Program of Higher Education (grant no. 20040005013) and the Beijing City Excellent Person Culture Fund (20061D0501500192). We also thank Dr. Yaoqi Zhou and Dr. Zhiping Weng for their helpful discussion.

References

[1] A. Tovchigrechko, C.A. Well, I.A. Vakser, Docking of protein models, Protein Sci. 11 (2002) 1888–1896.
[2] M.R. Chance, A.R. Bresnick, S.K. Burley, J.S. Jiang, C.D. Lima, A. Sali, S.C. Almo, J.B. Bonanno, J.A. Buglino, S. Boulton, H. Chen, N. Eswar, G. He, R. Huang, V. Ilyin, L. McMahan, U. Pieper, S. Ray, M. Vidal, L.K. Wang, Structural genomics: a pipeline for providing structures for the biologist, Protein Sci. 11 (2002) 723–738.
[3] A. Sali, R. Glaeser, T. Earnest, W. Baumeister, From words to literature in structural proteomics, Nature 422 (2003) 216–225.
[4] J. Janin, M. Levitt, Theory and simulation accuracy and reliability in modelling proteins and complexes, Curr. Opin. Struct. Biol. 16 (2006) 1–3.
[5] E. Katchalski-Katzir, I. Shariv, M.A. Eisenstein, C. Friesem, Aflalo, I. Vakser, Molecular surface recognition: determination of geometric fit between protein and their ligands by correlation techniques, Proc. Natl. Acad. Sci. U. S. A. 89 (1992) 2195–2199.
[6] P.H. Walls, M.J. Sternberg, New algorithm to model protein–protein recognition based on surface complementarity: applications to antibody–antigen docking, J. Mol. Biol. 228 (1992) 277–297.
[7] R. Rosenfeld, S. Vajda, C. DeLisi, Flexible docking and design, Annu. Rev. Biophys. Biomol. Struct. 24 (1995) 677–700.
[8] I.A. Vakser, Protein docking for low-resolution structures, Protein Eng. 8 (1995) 371–377.
[9] B. Sandak, R. Nussinov, H.J. Wolfson, A method for biomolecular structural recognition and docking allowing conformational flexibility, J. Comput. Biol. 5 (1998) 631–654.
[10] E. Althaus, O. Kohlbacher, H.P. Lenhof, P. Muller, A combinatorial approach to protein docking with flexible side chains, J. Comput. Biol. 9 (2000) 597–612.
[11] O. Kohlbacher, A. Burchardt, A. Moll, A. Hildebrandt, P. Bayer, H.-P. Lenhof, Structure prediction of protein complexes by a NMR-based protein docking algorithm, J. Biomol. NMR 20 (2001) 15–21.
[12] I. Halperin, B. Ma, H. Wolfson, R. Nussinov, Principles of docking: an overview of search algorithms and a guide to scoring functions, Proteins: Struct. Funct. Biol. 47 (2002) 409–443.
[13] B.J. McConkey, V. Sobolev, M. Edelman, Discrimination of native protein structures using atom-atom contact scoring, Proc. Natl. Acad. Sci. U. S. A. 83 (2002) 845–856.
[14] D.M. Lorber, M.K. Udo, B.K. Shoichet, Protein–protein docking with multiple residue conformations and residue substitutions, Protein Sci. 11 (2002) 1393–1408.
[15] J.J. Gray, S.E. Moughan, C. Wang, O. Schueler-Furman, B. Kuhlman, C.A. Rohl, D. Baker, Protein–protein docking with simultaneous optimization of rigid body displacement and side chain conformations, J. Mol. Biol. 331 (2003) 281–299.
[16] R. Chen, L. Li, Z. Weng, ZDOCK: an initial-stage protein-docking algorithm, Proteins: Struct. Funct. Biol. 52 (2003) 80–87.
[17] J. Fernandez-Recio, M. Totrov, R. Abagyan, ICM-DISCO docking by global energy optimization with fully flexible side-chains, Proteins: Struct. Funct. Biol. 52 (2003) 113–117.
[18] C. Dominguez, R. Boelens, A.M. Bonvin, HADDOCK: a protein–protein docking approach based on biochemical or biophysical information, J. Am. Chem. Soc. 125 (2003) 1731–1737.
[19] X.H. Ma, C.H. Li, L.Z. Shen, X.Q. Gong, W.Z. Chen, C.X. Wang, Biologically enhanced sampling geometric docking and backbone flexibility treatment with multiconformational superposition, Proteins: Struct. Funct. Biol. 60 (2005) 319–323.
[20] H.A. Gabb, R.M. Jackson, M.J. Sternberg, Modelling protein docking using shape complementarity, electrostatics and biochemical information, J. Mol. Biol. 272 (1997) 106–120.
[21] A. Heifetz, E. Katchalski-Katzir, M. Eisenstein, Electrostatics in protein–protein docking, Protein Sci. 11 (2002) 571–587.
[22] C. Zhang, G. Vasmatzis, J.L. Cornette, C. Delisi, Determination of atomic desolvation energies from the structures of crystallized proteins, J. Mol. Biol. 267 (1997) 707–726.
[23] G. Moont, H.A. Gabb, M.J.E. Sternberg, Use of pair potentials across protein interfaces in screening predicted docked complexes, Proteins: Struct. Funct. Biol. 35 (1999) 364–373.
[24] S. Liu, C. Zhang, H. Zhou, Y. Zhou, A physical reference state unifies the structure-derived potential of mean force for protein folding and binding, Proteins: Struct. Funct. Biol. 56 (2004) 93–101.
[25] J.J. Gray, S.E. Moughan, T. Kortemme, O. Schueler-Furman, K.M.S. Misura, A.V. Morozov, D. Baker, Protein–protein docking predictions for the Capri experiment, Proteins: Struct. Funct. Biol. 52 (2003) 118–122.
[26] S. Jones, J.M. Thornton, Principles of protein–protein interactions, Proc. Natl. Acad. Sci. U. S. A. 93 (1996) 113–120.
[27] L. Lo Conte, C. Chothia, J. Janin, The atomic structure of protein–protein recognition sites, J. Mol. Biol. 285 (1998) 2177–2198.
[28] M.J. Betts, M.J.E. Sternberg, An analysis of conformational changes on protein–protein association: implications for predictive docking, Protein Eng. 12 (1999) 271–283.
[29] R.M. Jackson, Comparison of protein–protein interactions in serine protease-inhibitor and antibody–antigen complexes: implications for the protein docking problem, Protein Sci. 8 (1999) 603–613.
[30] C.H. Li, X.H. Ma, W.Z. Chen, C.X. Wang, A protein–protein docking algorithm dependent on the type of the complexes, Protein Eng. 16 (2003) 256–269.
[31] R. Chen, J. Mintseris, J. Janin, Z. Weng, A protein–protein docking benchmark, Proteins: Struct. Funct. Biol. 52 (2003) 88–91.
[32] J. Mintseris, K. Wiehe, B. Pierce, R. Anderson, R. Chen, J. Janin, Z. Weng, Protein–protein docking benchmark 2.0: an update, Proteins: Struct. Funct. Biol. 60 (2005) 214–216.
[33] J. Fernández-Recio, M. Totrov, R. Abagyan, Identification of protein–protein interaction sites from docking energy landscapes, J. Mol. Biol. 335 (2004) 843–865.

[34] B. Kuhlman, D. Baker, Native protein sequences are close to optimal for their structures, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 10383–10388.
[35] B.R. Brooks, R.F. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus, Charmm: a program for macromolecular energy, minimization, and dynamics calculations, J. Comp. Chem. 4 (1983) 187–217.
[36] J. Murphy, D. Gatchell, J. Prasad, S. Vajda, Combination of scoring functions improves discrimination in protein–protein docking, Proteins: Struct. Funct. Biol. 3 (2003) 840–854.
[37] S.J. Hubbard, J.M. Thornton, NACCESS, Computer Program Department of Biochemistry and Molecular Biology, University College London, 1993.