ELSEVIER

Biophysical Chemistry 134 (2008) 111 – 118

Biophysical Chemistry

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

Amino acid network and its scoring application in protein–protein docking

Shan Chang, Xiong Jiao, Chun-hua Li, Xin-qi Gong, Wei-zu Chen, Cun-xin Wang *

College of Life Science and Bioengineering, Beijing University of Technology, Beijing 100022, China

Received 17 October 2007; received in revised form 4 December 2007; accepted 11 December 2007
Available online 26 December 2007

Abstract

Protein–protein complex, composed of hydrophobic and hydrophilic residues, can be divided into hydrophobic and hydrophilic amino acid network structures respectively. In this paper, we are interested in analyzing these two different types of networks and find that these networks are of small-world properties. Due to the characteristic complementarity of the complex interfaces, protein–protein docking can be viewed as a particular network rewiring. These networks of correct docked complex conformations have much more increase of the degree values and decay of the clustering coefficients than those of the incorrect ones. Therefore, two scoring terms based on the network parameters are proposed, in which the geometric complementarity, hydrophobic–hydrophobic and polar–polar interactions are taken into account. Compared with a two-term energy function, a simple scoring function HPNet which includes the two network-based scoring terms shows advantages in two aspects, not relying on energy considerations and better discrimination. Furthermore, combing the network-based scoring terms with some other energy terms, a new multi-term scoring function HPNet-combine can also make some improvements to the scoring function of RosettaDock.

Keywords: Hydrophobic; Hydrophilic; Amino acid network; Scoring function

1. Introduction

Molecular recognition plays a pivotal role in molecular biology. Molecular docking technology is an approach on protein–protein interaction and recognition. In the early days of docking study, the geometric matching has played an important role in determining the structure of a complex. Several approaches for scoring the results are based on geometric complementarity [1–3]. Later, hydrophobic–hydrophobic and polar–polar interactions between molecules have been considered in protein–protein docking. In the previous studies, it has been found that protein interfaces are generally consisted of a hydrophobic core surrounded by a more hydrophilic rim region [4–6]. In several studies, the contribution of hydrophobic–hydrophobic and polar–polar interactions has been taken into account in docking by combining surface complementarity with hydrophobic complementarity [7,8] or an electrostatic filter [9,10], with statistical potential methods such as residue potential scores or elaborate free energy evaluations [11–13].

Several real-world networks, such as the World Wide Web (WWW) [14,15], scientific collaborations networks [16] and the citation distribution of publications [17], show the similar topological properties. It is shown that they have large values of the clustering coefficient and small values of the characteristic path length. Watts and Strogatz [18] called them the small-world networks. Recently, the approach of small-world networks has become a powerful tool to study protein structure and function. Some application studies have focused on monomer proteins, such as the identification of key residues in protein folding [19] or functional residues in protein structure [20], the correlation between the average shortest path lengths and the residue fluctuations [21] and the relationship between the protein topological properties and their kinetic ability to fold [22–24]. Additionally, some groups have studied protein–protein complexes to identify key residues or hot spots in protein–protein interactions [25–28], to reveals the ligand binding network-bridging effects [29] and to analyze the flexible protein binding mechanisms [30].

The investigations of amino acid network have provided systematic and deeper understanding for the structure of proteins.

* Corresponding author. Tel.: +86 10 67392724; fax: +86 10 67392837.
E-mail address: cxwang@bjut.edu.cn (C. Wang).

0301-4622/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2007.12.005

In the docking process, many complex structures will be produced, and the few correct docked complex conformations have different structures from those incorrect ones. Furthermore, different complex structures will lead to different topologies of amino acid network. Therefore, the network representation of proteins could give some useful information for protein–protein docking. It has been found that hydrophobic–hydrophobic and polar–polar interactions are very important for protein–protein docking, but polar-hydrophobic interactions are not obvious [8]. Thus, in this work, we are only interested in analyzing the hydrophobic and hydrophilic amino acid networks of protein–protein complex separately, and quantitatively measuring the extent of hydrophobic complementarity and polar–polar interactions. Two scoring terms based on the network parameters are proposed in the present work. These scoring terms don't rely on energy considerations, but just need to measure the network parameters. Through these network parameters, geometric complementarity, hydrophobic–hydrophobic and polar–polar effects can be considered in the scoring terms. Then, these network-based scoring terms can be used for protein–protein docking.

2. Systems and methods

2.1. Docking datasets and decoys

A dataset of 43 dimer complexes was selected from the Benchmark 2.0 [31]. Only the single-chain monomer structures were selected to view the rewiring occurring in the system of two small-world networks. As a result, the new dataset consists of 18 enzyme/inhibitor cases, 1 antibody/antigen case and 24 others cases. The antibody/antigen case is the Target 06 (1KXQ) of CAPRI Round 2 (http://capri.ebi.ac.uk). The amino acid numbers of these complexes are from 126 to 915. To assess the accuracy of the scoring function, for each case, 1000 structures from bound and unbound docking were generated by the RosettaDock1.0 program [32], respectively.

2.2. Definition of surface residues, core residues and interface residues

The solvent accessible surface area (ASA) is used to define the three types of residues. It was calculated by the NACCESS program [33]. A residue is defined to be a surface residue (a residue on a protein surface) if its relative ASA is at least 25% of its nominal maximum area (the overall surface area of the residue that can be contacted by solvent), otherwise it is defined to be a core residue. A surface residue is defined to an interface residue if its ASA in the complex is less than that in the monomer by at least 1 $Å^2$ [34].

2.3. Networks of protein complexes and network parameters

The 20 kinds of amino acids are divided into two classes [35], in which the hydrophobic residues include Ile, Leu, Val, Phe, Met, Trp, Cys, Tyr, Pro and Ala, and the hydrophilic ones are Gly, Lys, Thr, Ser, Gln, Asn, Glu, Asp, Arg and His. Then, the protein complex structure is modeled as two undirected graphs. One is a hydrophobic amino acid network, in which hydrophobic residues being the vertices and atom contacts between them the edges. The other is a hydrophilic amino acid network, in which hydrophilic residues being the vertices and atom contacts between them the edges. Atom contacts are defined when any two atoms from two different residues are within 5.0 Å [23]. Thus, the adjacency matrix A is given by

$$A_{ij} = \begin{cases} 1 & \text{residue } i \text{ have atom contact with residue } j \\ 0 & \text{else} \end{cases} \quad (1)$$

The degree of any vertex i is given by

$$K_i = \sum_{j=1}^{N_v} A_{ij} \quad (2)$$

where A is the adjacency matrix of the undirected graph and N_v is the number of vertices. The average degree of the network can be written as

$$K = \frac{1}{N_v}\sum_i K_i \quad (3)$$

where N_v is the number of vertices.

The clustering coefficient of any vertex i is the ratio between the total number of links actually connecting its neighbours and the total number of all possible links between these neighbours. It is given by

$$C_i = \frac{n_i}{K_i(K_i-1)/2} \quad (4)$$

where K_i is the number of neighbors of the vertex i as defined in Eq. (2), and n_i is the actual number of edges between the neighbors of i. The clustering coefficient of network C is the average of C_i over all vertices. It is calculated as following

$$C = \frac{1}{N_v}\sum_i C_i \quad (5)$$

where N_v is the number of vertices.

The characteristic path length L is the average minimal distance between all pairs of vertices in the graph. It is calculated as following

$$L = \frac{1}{N_p}\sum_{j>i} l_{ij} \quad (6)$$

where N_p represents the number of pairs of vertices of the graph, and l_{ij} is the minimal path between vertices i and j [19].

2.4. Scoring functions and assessment criteria

The two-term energy-based function (EBscore) developed by our group [36] is calculated as

$$S_{\text{EBscore}} = \Delta G_{\text{des}} + \Delta E_{\text{elec}} \quad (7)$$

where the desolvation term ΔG_{des} is based on the atomic contact energy (ACE) model [37]

$$\Delta G_{\mathrm{des}} = \sum_i \sum_j e_{ij} \tag{8}$$

where e_{ij} denotes the atomic contact energy between atoms i and j, and the sum is taken over all atom pairs less than 6 Å apart. The electrostatic term ΔE_{elec} is calculated by the Coulombic potential function with a distance-dependent dielectric $\varepsilon = 4r$. The partial charges are from the CHARMM force field [38].

The two network-based scoring terms can be given by

$$S_{\mathrm{hnet}} = -\frac{K_h}{C_h} \tag{9}$$

and

$$S_{\mathrm{pnet}} = -\frac{K_p}{C_p} \tag{10}$$

where K_h and C_h are the average degree and the clustering coefficient of hydrophobic network, respectively. K_p and C_p are the average degree and the clustering coefficient of hydrophilic network, respectively. The hydrophobic–hydrophobic interactions can be evaluated by S_{hnet} and the term of polar–polar interactions is written as S_{pnet}. Then, a simple network-based scoring function HPNet can be composed of these two terms, which are similar to those of EBscore. It is calculated as

$$S_{\mathrm{hpnet}} = S_{\mathrm{hnet}} + S_{\mathrm{pnet}} \tag{11}$$

In order to do some comparisons with the general combined scoring functions, we combined the network-based scoring terms with other energy terms of RosettaDock [32] and made a new combined scoring function HPNet-combine. The logistic regression was used to determine weights that could maximally separate the good decoys from the others. Decoys were designated as "good" if the L_rmsds of them were less than 4.0 Å. The training set included 22 bound docking targets whose amino acid numbers are from 300 to 500. The other 21 bound docking results and all 43 unbound docking results are used to test the scoring performance. The regression results are shown in Table 1. The significance of each term in the scoring function is attested by its weight. Besides, the z-value of the assignment of the weight shows which terms aid in discriminating. Firstly, the RosettaDock energy terms and HPNet terms were all used to regression. Then, it was found that the z-values of short-range repulsive and short-range attractive are not high, so we eliminate the two terms and refit the weight. Then the scoring function HPNet-combine is a linear combination of these terms:

$$\begin{aligned} S_{\mathrm{hpnet-combine}} = {} & w_{\mathrm{rep}} S_{\mathrm{rep}} + w_{\mathrm{atr}} S_{\mathrm{atr}} + w_{\mathrm{sol}} S_{\mathrm{sol}} + w_{\mathrm{hbsc}} S_{\mathrm{hbsc}} \\ & + w_{\mathrm{hbbb}} S_{\mathrm{hbbb}} + w_{\mathrm{dun}} S_{\mathrm{dun}} + w_{\mathrm{pair}} S_{\mathrm{pair}} \\ & + w_{\mathrm{sasa}} S_{\mathrm{sasa}} + w_{\mathrm{elec}}^{\mathrm{lr-rep}} S_{\mathrm{elec}}^{\mathrm{lr-rep}} + w_{\mathrm{elec}}^{\mathrm{lr-atr}} S_{\mathrm{elec}}^{\mathrm{lr-atr}} \\ & + w_{\mathrm{hnet}} S_{\mathrm{hnet}} + w_{\mathrm{pnet}} S_{\mathrm{pnet}} \end{aligned} \tag{12}$$

where S_{rep} is a repulsive van der Waals score, S_{atr} is an attractive van der Waals score, S_{sol} is an implicit solvation score, S_{hbsc} is a side-chain hydrogen bonding score, S_{hbbb} is a backbone hydrogen bonding score, S_{dun} is a rotamer probability term, S_{pair} is a residue–residue pair probability term, S_{sasa} is a surface area-based solvation term, $S_{\mathrm{elec}}^{\mathrm{lr-rep}}$ is a long-range electrostatic repulsive term, $S_{\mathrm{elec}}^{\mathrm{lr-atr}}$ is a long-range electrostatic attractive term, S_{hnet} is the hydrophobic network term and S_{pnet} the polar network term. HPNet-combine has the same number of terms with the scoring function of RosettaDock and is suitable to compare with it.

Table 1
Weights used in the scoring function HPNet-combine

Score	Regression 1		Regression 2	
	Weight	Z-value	Weight	Z-value
Repulsive van der Waals	0.018	12.287	0.017	12.195
Attractive van der Waals	0.137	18.431	0.137	18.494
Gaussian solvent-exclusion	0.080	7.981	0.083	8.236
Hydrogen bonding (SC[a])	0.302	24.178	0.303	24.248
Hydrogen bonding (BB[b])	0.302	13.712	0.303	13.756
Rotamer probability	0.091	14.056	0.091	14.147
Residue pair probability	0.193	7.191	0.200	7.765
Surface area solvation	0.169	15.613	0.172	15.869
Simple electrostatics				
Short-range repulsive	−0.001	−0.258	–	–
Short-range attractive	0.042	2.465	–	–
Long-range repulsive	0.048	4.884	0.046	4.68
Long-range attractive	0.118	7.525	0.126	8.182
HPNet terms				
Hydrophobic network term	1.776	7.469	1.720	7.268
Polar network term	1.839	7.231	1.812	7.157

'–' indicates that the term is not included in regression.
[a] SC indicates side-chain.
[b] BB indicates backbone.

To assess the quality of the scoring, some values are used to evaluate the discriminative ability of the scoring functions, such as the correlation coefficients between the scoring values and L_RMSD, L_RMSD of the first rank, rank of the first hit and number of hits in top 10 scores. L_RMSD is computed over backbone atoms (N, C, CA, O) of the ligands after the receptors of the decoy are superimposed onto the native structure. The hit structure (or the near-native structure) is defined as the one with L_RMSD ≤ 4.0 Å for the bound docking structures and L_RMSD ≤ 5.0 Å for the unbound docking structures. In addition, it is very important for scoring function to find the higher accuracy structures. Therefore, the results with L_RMSD ≤ 2 Å are also analyzed in this work.

3. Results and discussion

3.1. Analysis of network properties

The network parameters of the 43 biologically diverse protein complexes are analyzed and the results of them are in agreement with previous studies [39]. Both the hydrophobic and hydrophilic networks of the protein complexes have large clustering coefficients and small characteristic path lengths. They exhibit the small-world network properties when compared with the random and regular graphs with the same number of vertices and average number of neighbors. The 86 monomer structures also

Table 2
The average values of the degree, the clustering coefficient and the characteristic path length for the amino acid networks of protein complexes and monomers

	K	C	L	Random network		Regular network	
				C_{rand}	L_{rand}	C_{reg}	L_{reg}
Hydrophobic network of complexes	6.72±0.39	0.488±0.018	5.68±1.31	0.040±0.019	2.74±0.21	0.618±0.009	15.25±5.89
Hydrophilic network of complexes	5.00±0.39	0.497±0.020	6.94±1.06	0.026±0.011	3.33±0.23	0.561±0.020	22.87±8.29
Hydrophobic network of monomers	6.37±0.67	0.503±0.037	4.01±1.26	0.089±0.057	2.40±0.29	0.608±0.021	8.43±5.14
Hydrophilic network of monomers	4.68±0.58	0.515±0.042	5.55±1.40	0.057±0.036	2.98±0.34	0.540±0.043	12.73±7.51

The clustering coefficients C and the characteristic path lengths L of random and regular graphs can be calculated as follows [19]

$$C_{rand} \sim \frac{K}{N}$$

$$L_{rand} \sim \frac{\ln N}{\ln K}$$

$$C_{reg} \sim \frac{3(K-2)}{4(K-1)}$$

$$L_{reg} \sim \frac{N(N+K-2)}{2K(N-1)}.$$

exhibit the same small-world properties as the complex structures (see Table 2).

Fig. 1 shows the degree value distributions of the core, interface and surface residues for amino acid networks of the protein–protein complex. Due to the rewiring of the monomer networks, the degree values of interface residues are higher than those of the surface residues in both of the hydrophobic and hydrophilic networks. Interestingly, this rewiring can also be seen from Table 2, in which both of the hydrophobic and hydrophilic complex networks have higher degree values than those of the monomer networks. In the hydrophobic network, the degree value distribution of the overall residues leans to the distribution of interface and core residues. On the opposite, in the hydrophilic network, the degree value distribution of the overall residues leans to the surface residues. This can be explained by the hydrophobic effect that burying of hydrophobic residues in protein core and exposure hydrophilic in protein surface can bring a larger entropy gain than burying of hydrophilic portions of the surface. Therefore, protein cores and interfaces are often made up of hydrophobic residues, whereas hydrophilic residues are more easily found in the surfaces. As shown in Fig. 1, it is found that different classes of interfaces have different network features. For the types of enzyme/inhibitor and others, the degree values of interface residues are both high in the hydrophobic and hydrophilic networks. It shows that the hydrophobic and hydrophilic interactions are both important for enzyme/inhibitor and others. While for the antibody/antigen, the interface degree values of the hydrophobic networks is lower and the interface degree values of the

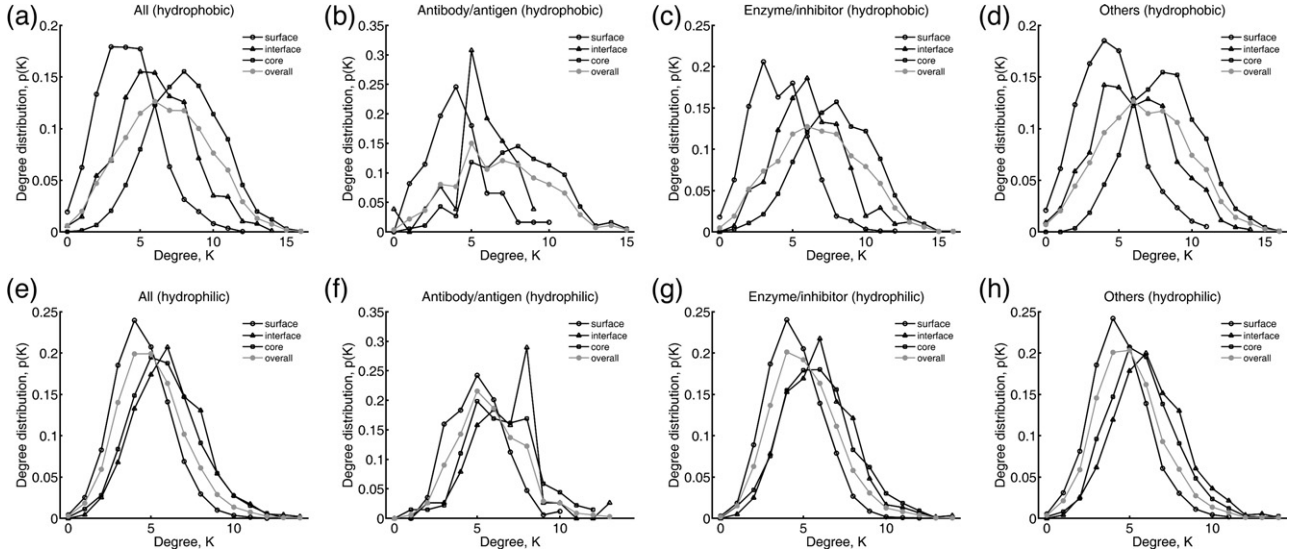

Fig. 1. The degree value distribution P(K) of the protein–protein complex amino acid networks. (a) Hydrophobic networks of all types. (b) Hydrophobic networks of antibody/antigen. (c) Hydrophobic networks of enzyme/inhibitor. (d) Hydrophobic networks of others. (e) Hydrophilic networks of all types. (f) Hydrophilic networks of antibody/antigen. (g) Hydrophilic networks of enzyme/inhibitor. (h) Hydrophilic networks of others.

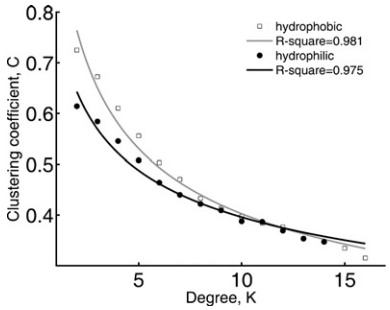

Fig. 2. The vertex clustering coefficient $C(K)$ as a function of vertex degree K for the amino acid network. Both the clustering coefficients of the hydrophobic and the hydrophilic networks exhibit the power-law decay as a function of K. The best fit curves are shown by lines.

hydrophilic networks is higher than those of the enzyme/inhibitor and others. It reflects the generally polar nature of the antigen's paratope. These results are consistent with previous studies [40].

As shown in Fig. 2, the vertex clustering coefficients C exhibit the power-law decay with increasing of the vertex degree K for both the hydrophobic and the hydrophilic networks, where $C(K)=aK^{-\beta}$. The scaling coefficient β for the hydrophobic networks is 0.399; whereas β for the hydrophilic networks is 0.302. This can be explained that due to spatial constrains imposed by neighboring residues within the cluster, the vertices have less number of interactions between them [41]. Therefore, as the rewiring of the monomer networks happens, the degree values of vertex increase whereas the clustering coefficients decrease.

Protein–protein docking is required to satisfy the hydrophobic complementarity and the polar–polar specific recognition. Thus, if docked at the correct position, the complex structure needs to have a good geometric complementarity and characteristic matching on the interface. Due to the complex interfaces being of characteristic complementarity, the protein–protein docking can be viewed as the particular network rewiring in the system of the two monomers (see Fig. 3). But it is hard for incorrect docked structures to satisfy these requirements. Therefore, the rewired structure of incorrect docked have different network properties from the correct ones. First, as shown in Fig. 3(a) and (b), their geometric complements are not good. Second, for Fig. 3(a) and (c), the hydrophobic and hydrophilic characteristics of interface residues are not matching. From Fig. 3(d), it can be observed that the hydrophobic and hydrophilic networks of correct docked structures will have much more increase of the degree values and decay of the clustering coefficients than that of the incorrect ones. Therefore, the scoring terms is designed to differentiate the two incorrect docking types from the correct docking results (see Systems and methods).

3.2. Scoring functions performance and comparison

Two comparisons are made to show the efficiency of the network parameters in the discrimination of protein–protein docking. One comparison is between the two-term energy function EBscore and the network-based scoring function HPNet. EBscore and HPNet are both composed of two terms, while the general combined scoring functions often include more than three terms. Therefore, here we choose EBscore to be compared with HPNet. Another comparison is between the scoring function of RosettaDock and HPNet-combine. RosettaDock has a multi-term energy function in protein–protein docking. It is compared with the HPNet-combine which includes the same number of terms. The comparison results are summarized in Table 3. The detailed information is listed in the supplementary material Tables 4–7.

First, we compared the discriminative abilities of EBscore and HPNet. For the bound docking results in Table 4, 15 out of the 18 enzyme/inhibitor complex cases and 16 out of the 24 others complex cases, HPNet gives better or the same ranking places for the near-native docked structures than EBscore does. For the 16 enzyme/inhibitor and the 19 others cases, the correlation coefficients between L_RMSD and HPNet are higher than those between L_RMSD and EBscore. Using EBscore, there are only 6 enzyme/inhibitor cases and 2 others cases with the first higher accuracy structure ranked within the top 10, 7 enzyme/inhibitor cases and 3 others cases with the first hit ranked within the top 10, and 3 enzyme/inhibitor cases and 3 others cases with a hit ranked first. While using HPNet, there are 8 enzyme/inhibitor cases and 9 others cases with the first higher accuracy structure ranked within the top 10, 9 enzyme/inhibitor cases and 12 others cases with the first hit ranked within the top 10, and 4 enzyme/inhibitor cases and 5 others cases with a hit ranked first. The total hits number within the top 10 by using EBscore is 54, while for HPNet it is increased to127.

For the unbound docking results in Table 5, L_RMSD of the first rank is bigger than that of the bound docked structures for most cases. Therefore, we define the hit structure of unbound docking is the one with L_RMSD ≤ 5.0 Å. For the cases of the others type 1GHQ, 1H1V and 1IBR, no docked structure with

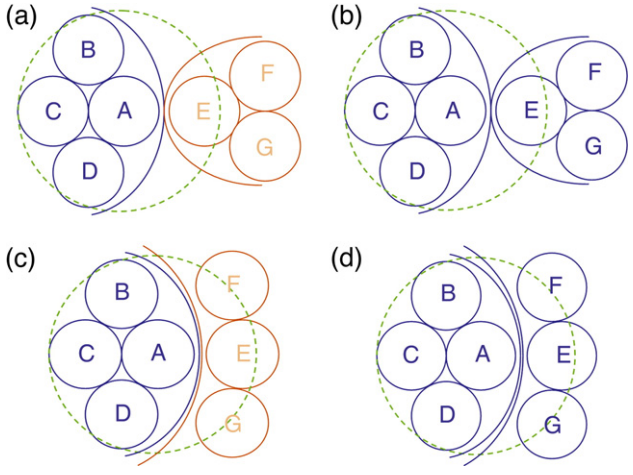

Fig. 3. The schematic diagram of the rewiring on the two monomer amino acid networks. The large dashed circle represents the atom contact distance from the amino acid 'A'. Any amino acid within the large circle should have interactions with the amino acid 'A'. The blue color amino acid has a different characteristic from the orange color amino acid. (a) The docked result is of bad shape complementarity and bad characteristic matching. (b) The docked result is of bad shape complementarity and good characteristic matching. (c) The docked result is of good shape complementarity and bad characteristic matching. (d) The docked result is of good shape complementarity and good characteristic matching.

Table 3
The discriminative ability comparison of scoring functions

	Antibody/antigen			Enzyme/inhibitor			Others			Total hits [d]
	Higher accuracy [a]	Hit [b]	Hit ranked 1st [c]	Higher accuracy	Hit	Hit ranked 1st	Higher accuracy	Hit	Hit ranked 1st	
Bound										
EBscore	0	0	0	6	7	3	2	3	3	54
HPNet	1	1	0	8	9	4	9	12	5	127
RosettaDock	1	1	1	11	14	9	12	17	13	101/125/226 [e]
HPNet-combine	1	1	1	13	16	9	18	20	15	127/141/268
Unbound										
EBscore	0	0	0	0	2	1	0	1	1	11
HPNet	1	1	0	2	6	0	1	3	1	33
RosettaDock	1	1	1	4	12	8	2	7	3	114
HPNet-combine	1	1	1	6	14	8	4	8	2	131

[a] The number of cases with the first higher accuracy structure (L_RMSD ≤ 2 Å) ranked within the top 10.
[b] The number of cases with the first hit (L_RMSD ≤ 4 Å for bound or/and L_RMSD ≤ 5 Å for unbound docking structures) ranked within the top 10.
[c] The number of cases with a hit ranked first.
[d] The total number of hits can be found by the scoring function.
[e] The first number shows the hits number on the 22 training proteins. The second number shows the hits number on the 21 test proteins. The third number is the total hits number of the 43 proteins.

L_RMSD less than 5.0 Å is obtained at the sampling stage. Two cases (1H1V, 1IBR) are referred as difficult in Benchmark 2.0 [31]. These difficult cases have the large molecular conformational changes upon complex formation, which is difficult for sampling. Therefore, in the others cases of unbound docking results, 21 proteins have the near-native docked structures. For 15 out of the 18 enzyme/inhibitor complex cases and 15 out of the 21 others complex cases, HPNet gives better and higher ranking places for the near-native docked structures than EBscore does. For the 17 enzyme/inhibitor and the 18 others cases, the correlation coefficients between L_RMSD and HPNet are higher than those between L_RMSD and EBscore. Using EBscore, there is no any enzyme/inhibitor case or others case with the first higher accuracy structure ranked within the top 10, 2 enzyme/inhibitor cases and 1 others case with the first hit ranked the top 10, and 1 enzyme/inhibitor case and 1 others case with a hit ranked first. While for HPNet, there are 2 enzyme/inhibitor cases and 1 others case with the first higher accuracy structure ranked within the top 10, 6 enzyme/inhibitor cases and 3 others cases with the first hit ranked within the top 10, and 0 enzyme/inhibitor case and 1 others case with a hit ranked first. The total hits number within the top 10 by using EBscore is 11, while for HPNet it is increased to 33.

From the comparison, it is found that HPNet shows relatively better discriminative ability than EBscore. It is notable that the complex 1KXQ (Antibody Vhh Fragment/Amylase) [42] is a target of CAPRI Round 2. As to both the bound and unbound structures of this target, EBscore fails to obtain the near-native conformations, whereas HPNet can rank the near-native conformations within the top 10. These comparisons show that the network parameters can be a better choice for scoring terms. However, some failures, especially in the others unbound cases, have also been found for HPNet. From current studies [43,44], it is known that a good predictor of near-native structure should be the multi-term scoring function, and the different terms can supply different contributions for the scoring function. Since the HPNet is composed of only two terms, it is difficult for this scoring function to have better discrimination than the combined scoring functions including many energy terms. However, the scoring function HPNet is simple and relatively effective, and it implements the network tools in protein–protein scoring function for the first time. Meanwhile, if combined with the network terms of HPNet, the combined scoring functions could have some improvements.

Second, we compared the discriminative abilities of the multi-term scoring function of RosettaDock and HPNet-combine. For the bound docking results in Table 6, 14 out of the 18 enzyme/inhibitor complex cases and 24 out of the 24 others complex cases, HPNet-combine gives better or the same ranking places for the near-native docked structures than the scoring function of RosettaDock does. For the 16 enzyme/inhibitor and the 18 others cases, the correlation coefficients between L_RMSD and HPNet-combine are higher than those between L_RMSD and the scoring function of RosettaDock. Using the scoring function of RosettaDock, there are only 11 enzyme/inhibitor cases and 12 others cases with the first higher accuracy structure ranked within the top 10, 14 enzyme/inhibitor cases and 17 others cases with the first hit ranked within the top 10, and 9 enzyme/inhibitor cases and 13 others cases with a hit ranked first. While using HPNet-combine, there are 13 enzyme/inhibitor cases and 18 others cases with the first higher accuracy structure ranked within the top 10, 16 enzyme/inhibitor cases and 20 others cases with the first hit ranked within the top 10, and 9 enzyme/inhibitor cases and 15 others cases with a hit ranked first. The total hits number within the top 10 by using the scoring function of RosettaDock is 226, while for HPNet-combine it is increased to 268.

Since the regression is trained on bound structures, the unbound docking structures can be seen as the test set. The comparison of unbound docking results is shown in Table 7. For 12 out of the 18 enzyme/inhibitor complex cases and 17 out of the 21 others complex cases, the scoring function HPNet-combine gives better or the same ranking places for the near-native docked structures than the scoring function of RosettaDock does. For the 10 enzyme/inhibitor and the 14 others cases,

the correlation coefficients between L_RMSD and HPNet-combine are higher than those between L_RMSD and the scoring function of RosettaDock. Using the scoring function of RosettaDock, there are only 4 enzyme/inhibitor cases and 2 others cases with the first higher accuracy structure ranked within the top 10, 12 enzyme/inhibitor cases and 7 others cases with the first hit ranked the top 10, and 8 enzyme/inhibitor cases and 3 others cases with a hit ranked first. While for HPNet-combine, there are 6 enzyme/inhibitor cases and 4 others cases with the first higher accuracy structure ranked within the top 10, 14 enzyme/inhibitor cases and 8 others cases with the first hit ranked within the top 10, and 8 enzyme/inhibitor cases and 2 others cases with a hit ranked first. The total hits number within the top 10 by using the scoring function of RosettaDock is 114, while for HPNet-combine it is increased to 131.

From the comparison on the bound and unbound docking decoys, it is found that HPNet-combine shows relatively better than the scoring function of RosettaDock. However, for the regression, the function needs to show its credible extension by obtaining good performances on the test set. Therefore, we should pay more attention to the 21 test proteins and the 43 unbound decoys. As shown in Table 3, by using the scoring function of RosettaDock on the test proteins, the total hits number is 125, while for HPNet-combine it is 141. The number is increased 12.8%. Meanwhile, by using the scoring function of RosettaDock on unbound docking decoys, the total hits number is 114, while for HPNet-combine it is 131. The number is increased 14.9%. Thus, it is shown that HPNet-combine improves the scoring function of RosettaDock more than 10% in the discriminative ability. At the same time, we should note that their discrimination abilities are different for different classes of complexes. For the types of enzyme-inhibitor and others, HPNet-combine can find more hit structures than the scoring function of RosettaDock. While for the unbound docking decoys of antibody/antigen, the scoring function of RosettaDock finds 9 hit structures, which is one more than that of HPNet-combine (see Table 7). This can be explained by the fact that RosettaDock considered more electrostatic terms than HPNet-combine, which account better for the polar interaction of the antigen's paratope.

Although HPNet-combine has better performance with implementation of the new network parameters, the polar/hydrophobic complimentarity essence of the parameters may have some redundancies with the general scoring energy terms of RosettaDock, which will make the z-values of them relatively low (see Table 1). Additionally, protein residues are just divided into two types and both of the hydrophobic and hydrophilic networks are un-weighted in our network model. These simplifications would affect the z-values. If the weighted networks and other network parameters are considered in the scoring function, the z-values of network terms will be improved. This work is currently underway.

4. Conclusion

In this work, the hydrophobic and hydrophilic amino acid networks of the protein–protein complex are analyzed separately. Both of them show the character of small-world. Since the network tools can take account of the global topological characteristics of the protein–protein complex, this method can makes it easy to explain the physics principle of protein–protein interactions. By using the simple network model, the protein–protein docking can be viewed as the rewiring of the monomer networks. Thus, two network-based scoring terms are proposed for protein–protein docking. It is just composed of the elementary network parameters, the degree and clustering coefficient. The simple scoring function HPNet including the two network-based scoring terms does not rely on energy considerations and shows a better performance than the two-term energy function EBscore. Furthermore, these network-based scoring terms have also been used in conjunction with other scoring terms and the new multi-term scoring HPNet-combine is devised. It can improve the discrimination of the combined scoring function of RosettaDock more than 10%. This work might provide some insight into the future development of the scoring functions on protein–protein complex.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (10574009, 20773006) and Specialized Research Fund for the Doctoral Program of Higher Education (20040005013).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bpc.2007.12.005.

References

[1] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A.A. Friesem, C. Aflalo, I.A. Vakser, Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques, Proc. Natl. Acad. Sci. U. S. A. 89 (1992) 2195–2199.
[2] P.H. Walls, M.J.E. Sternberg, New algorithm to model protein–protein recognition based on surface complementarity. Applications to antibody–antigen docking, J. Mol. Biol. 228 (1992) 277–297.
[3] R. Norel, S.L. Lin, H.J. Wolfson, R. Nussinov, Shape complementarity at protein–protein interfaces, Biopolymers 34 (1994) 933–940.
[4] A.A. Bogan, K.S. Thorn, Anatomy of hot spots in protein interfaces, J. Mol. Biol. 280 (1998) 1–9.
[5] P. Chakrabarti, J. Janin, Dissecting protein–protein recognition sites, Proteins 47 (2002) 334–343.
[6] R.P. Bahadur, P. Chakrabarti, F. Rodier, J. Janin, Dissecting subunit interfaces in homodimeric proteins, Proteins 53 (2003) 708–719.
[7] F. Ackermann, G. Herrmann, S. Posch, G. Sagerer, Estimation and filtering of potential protein–protein docking positions, Bioinformatics 14 (1998) 196–205.
[8] A. Berchanski, B. Shapira, M. Eisenstein, Hydrophobic complementarity in protein–protein docking, Proteins 56 (2004) 130–142.
[9] R. Norel, F. Sheinerman, D. Petrey, B. Honig, Electrostatic contributions to protein–protein interactions: fast energetic filters for docking and their physical basis, Protein Sci. 10 (2001) 2147–2161.
[10] A. Heifetz, E. Katchalski-Katzir, M. Eisenstein, Electrostatics in protein–protein docking, Protein Sci. 11 (2002) 571–587.
[11] R.M. Jackson, H.A. Gabb, M.J.E. Sternberg, Rapid refinement of protein interfaces incorporating solvation: application to the docking problem, J. Mol. Biol. 276 (1998) 265–285.

[12] G. Moont, H.A. Gabb, M.J.E. Sternberg, Use of pair potentials across protein interfaces in screening predicted docked complexes, Proteins 35 (1999) 364–373.
[13] C.J. Camacho, D.W. Gatchell, S.R. Kimura, S. Vajda, Scoring docked conformations generated by rigid-body protein–protein docking, Proteins 40 (2000) 525–537.
[14] R. Albert, H. Jeong, A.L. Barabasi, Internet: diameter of the world-wide web, Nature 401 (1999) 130–131.
[15] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the web, Comput. Networks 33 (2000) 309–320.
[16] M.E.J. Newman, The structure of scientific collaboration networks, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 404–409.
[17] S. Redner, How popular is your paper? An empirical study of the citation distribution, Eur. Phys. J. B 4 (1998) 131–134.
[18] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (1998) 440–442.
[19] M. Vendruscolo, N.V. Dokholyan, E. Paci, M. Karplus, Small-world view of the amino acids that play a key role in protein folding, Phys. Rev. E 65 (2002) 061910.
[20] A. Shemesh, G. Amitai, E. Sitbon, M. Shklar, D. Netanely, I. Venger, S. Pietrokovski, Structural analysis of residue interaction graphs, ISMB/ECCB2004, 2004, pp. 22–23.
[21] A.R. Atilgan, P. Akan, C. Baysal, Small-world communication of residues and significance for protein dynamics, Biophys. J. 86 (2004) 85–91.
[22] N.V. Dokholyan, L. Li, F. Ding, E.I. Shakhnovich, Topological determinants of protein folding, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 8637–8641.
[23] L.H. Greene, V.A. Higman, Uncovering network systems within protein structures, J. Mol. Biol. 334 (2003) 781–791.
[24] X. Jiao, S. Chang, C.H. Li, W.Z. Chen, C.X. Wang, Construction and application of the weighted amino acid network based on energy, Phys. Rev. E 75 (2007) 051903.
[25] A. del Sol, H. Fujihashi, D. Amoros, R. Nussinov, Residues crucial for maintaining short paths in network communication mediate signaling in proteins, Mol. Syst. Biol. 2 (2006) 0019.
[26] A. del Sol, P. O'Meara, Small-world network approach to identify key residues in protein–protein interaction, Proteins 58 (2005) 672–682.
[27] A. del Sol, H. Fujihashi, P. O'Meara, Topology of small-world networks of protein–protein complex structures, Bioinformatics 21 (2005) 1311–1315.
[28] K.V. Brinda, S. Vishveshwara, Oligomeric protein structure networks: insights into protein–protein interactions, BMC Bioinformatics 6 (2005) 296–310.
[29] Z.J. Hu, D. Bowen, W.M. Southerland, A. del Sol, Y.P. Pan, R. Nussinov, B.Y. Ma, Ligand binding and circular permutation modify residue interaction network in DHFR, PLoS Comput. Biol. 3 (2007) e117.
[30] Y. Levy, S.S. Cho, J.N. Onuchic, P.G. Wolynes, A survey of flexible protein binding mechanisms and their transition states using native topology based energy landscapes, J. Mol. Biol. 346 (2005) 1121–1145.
[31] J. Mintseris, K. Wiehe, B. Pierce, R. Anderson, R. Chen, J. Janin, Z.P. Weng, Protein–protein docking benchmark 2.0: an update, Proteins 60 (2005) 214–216.
[32] J.J. Gray, S. Moughon, 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.
[33] S.J. Hubbard, J.M. Thornton, NACCESS [computer program].Department of Biochemistry and Molecular Biology, University College London, 1993.
[34] S. Jones, J.M. Thornton, Principles of protein–protein interactions, Proc. Natl. Acad. Sci. U. S. A. 93 (1996) 13–20.
[35] S.J. Sun, R. Brem, H.S. Chan, K.A. Dill, Designing amino acid sequences to fold with good hydrophobic cores, Protein Eng. 8 (1995) 1205–1213.
[36] C.H. Li, X.H. Ma, W.Z. Chen, C.X. Wang, A soft docking algorithm for predicting the structure of antibody–antigen complexes, Proteins 52 (2003) 47–50.
[37] 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.
[38] B.R. Brooks, R.E. 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.
[39] S. Kundu, Amino acid network within protein, Physica A 346 (2005) 104–109.
[40] 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.
[41] M. Aftabuddin, S. Kundu, Weighted and unweighted network of amino acids within protein, Physica A 369 (2006) 895–904.
[42] A. Desmyter, S. Spinelli, F. Payan, M. Lauwereys, L. Wyns, S. Muyldermans, C. Cambillau, Three camelid VHH domains in complex with porcine pancreatic alpha-amylase. Inhibition and versatility of binding topology, J. Biol. Chem. 277 (2002) 23645–23650.
[43] R. Mendez, R. Leplae, M.F. Lensink, S.J. Wodak, Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures, Proteins 60 (2005) 150–169.
[44] V. Mohan, A.C. Gibbs, M.D. Cummings, E.P. Jaeger, R.L. DesJarlais, Docking: successes and challenges, Curr. Pharm. Des. 11 (2005) 323–333.