Analysis of Protein Surface Patterns by Pocket Similarity Network

Zhi-Ping Liu[1,2], Ling-Yun Wu[1,*], Yong Wang[1], Xiang-Sun Zhang[1] and Luonan Chen[3,4,5,6]

[1]*Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China;* [2]*Graduate University of Chinese Academy of Sciences, Beijing 100049, China;* [3]*Institute of Systems Biology, Shanghai University, Shanghai 200444, China;* [4]*Osaka Sangyo University, Osaka 574-8530, Japan;* [5]*ERATO Aihara Complexity Modelling Project, JST, Tokyo 151-0064, Japan;* [6]*Institute of Industrial Science, The University of Tokyo, Tokyo 153-8505, Japan*

Abstract: In this work, in order to reveal protein surface patterns in a systems biology framework, we analyze the similarity among the surface cavities by investigating the features of the pocket similarity network such as the community structure, the small-world property, the scale-free characteristic, and the hubs.

Keywords: Protein surface pattern, pocket similarity, complex network, functional genomics, systems biology.

1. INTRODUCTION

Systems biology, in some sense, is the study of the interactions between the components of a biological system by a network concept, and of showing that how these interactions give rise to the function and behavior of the system [1,2]. Many complicated systems can be represented as networks of interactions among individual components. The network properties can characterize the whole system and its individual components [3-5] at the same time, thus are generally able to be applied to many disciplines. Examples include social networks (e.g. scientific collaboration networks), technological networks (e.g. the world-wide web and power grids), and biological networks (e.g. neural networks, cellular and metabolic networks) [6,7]. A network model often abstracts the components as nodes (vertices) and their relationships as edges (lines) in a graph, where the weights companying with nodes and/or edges represent the degree or constraint of the relationships [7]. With the increasing availability of large-scale high-throughput data, network-based methods have shed light on research of protein science, such as the protein-protein interactions, the domain-domain interactions, and the amino acid contacts within protein structures [5,8-10]. These investigations provided deep understanding of the evolution and diversity in protein universe.

In this paper we aim to reveal protein surface patterns or structural motifs by using the network concept to integrate the existing protein structure data. Since more and more protein 3D structures are available with the rapid progress of structural genomics, the challenge is how to extract useful and valuable knowledge on the biochemical and biological roles of proteins from the structural data [11,12]. It is well known that functions of a protein are mainly determined by its physical, chemical and geometric properties of structural surfaces, which are the places where a protein carries out its functions [13-15]. Recently, there are some studies [16-18] focusing on the analysis of the molecular surface, because a detailed characterization of the protein surface is a key element in understanding and predicting the binding preferences of diverse proteins [19-22]. Furthermore the specificity of the protein surface provides various active sites which are the targets of protein-protein interface [24-26]. Thus in this paper, we only consider the protein surface regions in a protein structure, such as pockets or clefts which provide specialized environment for biological activities. The underlying 3D shape and physicochemical texture in these regions facilitate functional interactions [19-21,23]. There are already some results showing that small ligands tend to bind proteins at large surface pockets and perform critical functions, such as binding [19,28].

With the protein surface regions (pockets) as the basic elements in this work, we establish links among them to form a network. Structural similarity between pairs of the surface patterns is a natural choice since it gives evidences of functional relationships between the proteins where they are located [27-30]. Moreover, analyzing the similarity among these surface regions from a systematic viewpoint not only will provide valuable hints to study the conserved functional surface patterns in evolution, but also can gain deep insights into the biochemical relationships between functions and structural motifs. In addition, the knowledge about surface motifs is also crucial for drug design and other bioengineering.

Specifically, in the constructed pocket similarity network, we define a node by a protein surface region such as a pocket, and define an edge by the similarity relationship between two pockets. Then, we explore the surface concavity patterns by analyzing the features of the pocket similarity network from the systematic viewpoint. The results show that the pocket similarity network contains not only the common features as other complex networks but also its own special characteristics. In this paper, the pocket similarity network is firstly constructed in a systematic manner by linking the pockets according to their similarity scores. Then we investigate the community structure architectures underlying the network. The small-world character and the scale-free distribution of the network are also presented. The physicochemical property of the hub pockets is discussed to under-

*Address correspondence to this author at the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China; Tel: +86-10-62616659; Fax: +86-10-62561963; E-mail: lywu@amt.ac.cn

0929-8665/08 $55.00+.00

stand the implications of the network architecture. Finally, the future research topics as well as the potential applications of the network are discussed.

2. METHODS

2.1. Constructing Pocket Similarity Network

Binding sites and active sites of proteins and DNAs are often associated with structural pockets and cavities. Generally, a pocket can be regarded as an empty concavity on a protein surface into which solvent can gain access, i.e. the concavities have the mouth openings connecting their interior with the outside bulk solution [20]. Fig. (**1**) gives an example of a pocket.

There are many definitions for protein pockets from various aspects. In this paper, we adopt the pocket definition in CASTp [31] and pvSOAR [32] databases. The CASTp server uses the weighted Delaunay triangulation and the alpha complex for shape measurements. It provides identification and measurements of surface accessible pockets as well as interior inaccessible cavities, for proteins and other molecules. It measures analytically the area and volume of each pocket and cavity, both in solvent accessible surface (SA, Richards' surface) and molecular surface (MS, Connolly's surface). Obviously, it can be straightforwardly extended to other existing pocket or cleft definitions.

We construct the similarity network to describe the global relationship among the surface pockets. To remove the redundancy in PDB [33], we use the proteins in PDBselect25 [34], in which the proteins have low sequence similarity($<$25%) and come from different protein families. We collect all the pockets of proteins in PDBselect25 from CASTp database (more than 78900). Each pocket is represented by a node. Two nodes are linked by an edge if their structural similarity is larger than a given threshold. The similarities among the pockets are measured by pvSOAR (pocket and void surfaces of amino acid residues) database, which detects surface similarities in protein structures. It allows a user to search a protein surface pattern derived from a pocket or a void against all known surface patterns from CASTp database. When querying one pocket in pvSOAR, it would hit some similar pockets within a given threshold. The pvSOAR database compares the pockets in CASTp [32] in an all-against-all way. Fig. (**2**) gives an example, where we use the loosest threshold, i.e. structural cRMSD (coordinate root mean square distance) p-value 0.9. Thus an edge in the pocket similarity network links two structurally similar pockets. The isolated nodes in the network are discarded, because they march no similar pockets by the given threshold in the whole pocket library.

2.2. Analyzing Network Properties

With the constructed similarity network, analysis of the network properties is carried out by using the concepts and ideas as in other complex networks [6]. To find out the unique features of the pocket similarity network, we use thresholds from 0.9 (loosest) to 0.1 (tightest) to filter out the subnetworks with different similarities. The special characteristics of the similarity networks such as community structure feature are expected to be identified by analyzing these subnetworks. More detailed analysis is conducted on the maximum connected component in each subnetwork. The properties of these maximum connected components such as the small-world feature and the scale-free distribution are detected. For the detailed definitions of the characteristics such as path length, clustering coefficient and other measurement of the complex networks, readers can refer to [3,4,6,8]. In addition, we also identify the hub pockets in the similarity network, which have relatively more connections with other similar pockets in the database. Analyzing the physicochemical features of them can elucidate more interesting properties of the network, and it provides more biological and evolutionary implications on functional motifs.

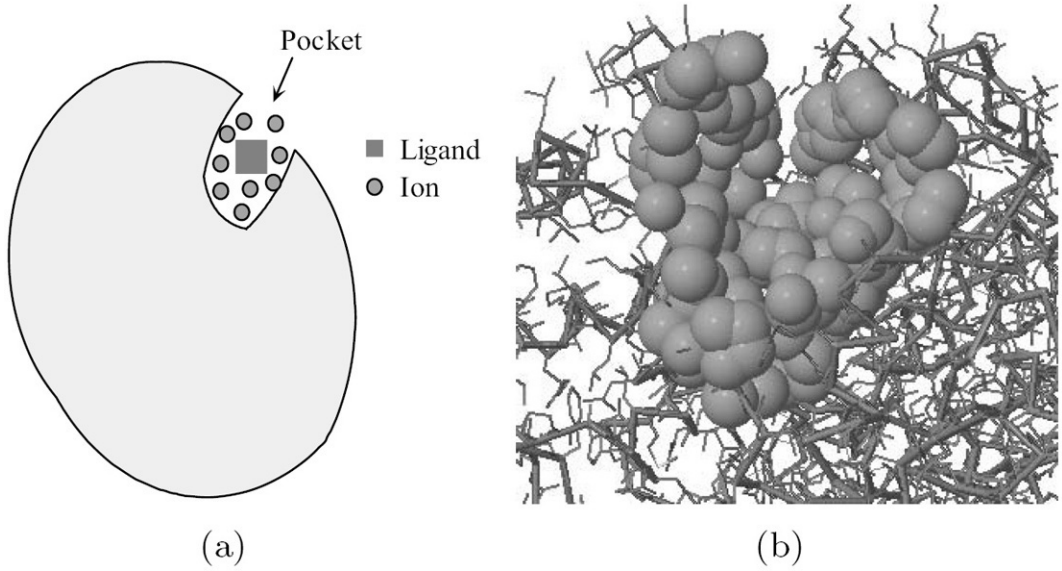

Figure 1. (a) An illustration of a pocket on a protein surface. (b) One true pocket (created from CASTp).

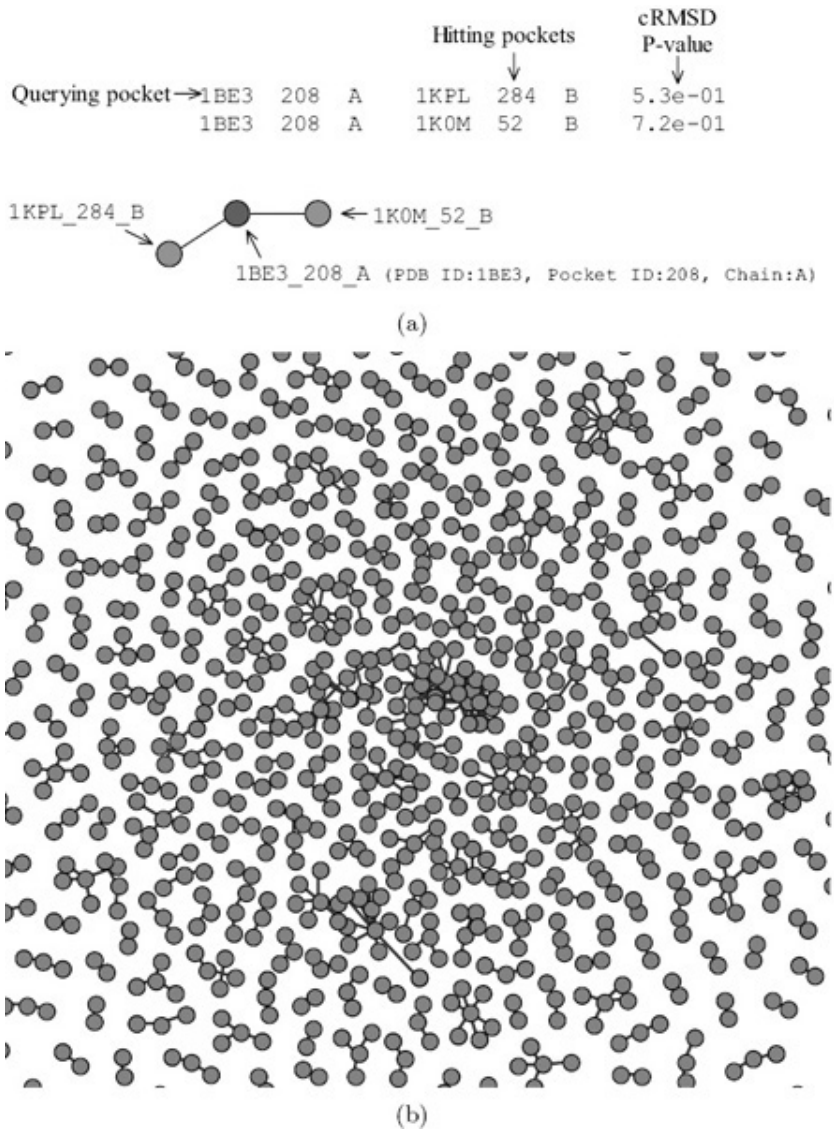

Figure 2. The illustration of the method to construct the pocket similarity network. (a) An example of the querying results. (b) A part of the pocket similarity network.

3. RESULTS

3.1. Community Structure of the Pocket Similarity Networks

Firstly we identified the physical architecture of the pocket similarity network by viewing it as a complex network and found the unique properties of the particular system. As shown in Fig. (**2**), the pocket similarity network tends to be sparse and possesses many clusters formed by small connected components spontaneously. This fact indicates that the pocket similarity network can be explored to analyze protein structures and functions based on its community structures. Another evidence is that a subnetwork with higher similarity is sparser than that with lower similarity. Table **1** shows the tendency.

As shown in Table **1**, the number of the connected components varies when different thresholds (i.e. different structural similarity degrees) are applied. It is obvious that when the similarity degree of the subnetwork is increased, the network becomes sparser and is divided into more connected components. In Fig. (**3**), we further recorded the percentages and the numbers of non-trivial connected components (at least two nodes) with different sizes. The statistical results show that the pocket similarity network possesses clear community structure.

The reasons for special architecture of the pocket similarity network lie in the properties of the structural similarity among protein surface patterns. Firstly, the similar pockets tend to cluster together by the particular similarity which is different from classical metric. The classical metric satisfies transitivity (e.g. if A is similar to B, and B is similar to C, then A is similar to C). The special structure in the pocket similarity network reveals the fact that similarity among pockets are not transitive, different from classical metric, although the other metric properties such as reflexivity (e.g. A is similar to itself), symmetry (e.g. A is similar to B, then

Table 1. The Number of Nodes, Edges and Connected Components of the Pocket Similarity Networks Constructed by Different Thresholds (cRMSD p-Value from 0.9 to 0.1). "Component" Means the Number of Non-Trivial Connected Components (i.e. at Least Two Nodes), and the Number in the Parentheses is the Number of the Discarded Nodes (i.e. Isolated Nodes) Compared with Connected Subnetworks

Threshold	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Node	5387	4907	4421	3957	3522	3048	2579	2018	1455
Edge	4943	4259	3681	3158	2704	2274	1854	1408	1002
Component	880	980 (480)	1016 (486)	1023 (464)	995 (435)	920 (474)	827 (469)	693 (561)	520 (563)

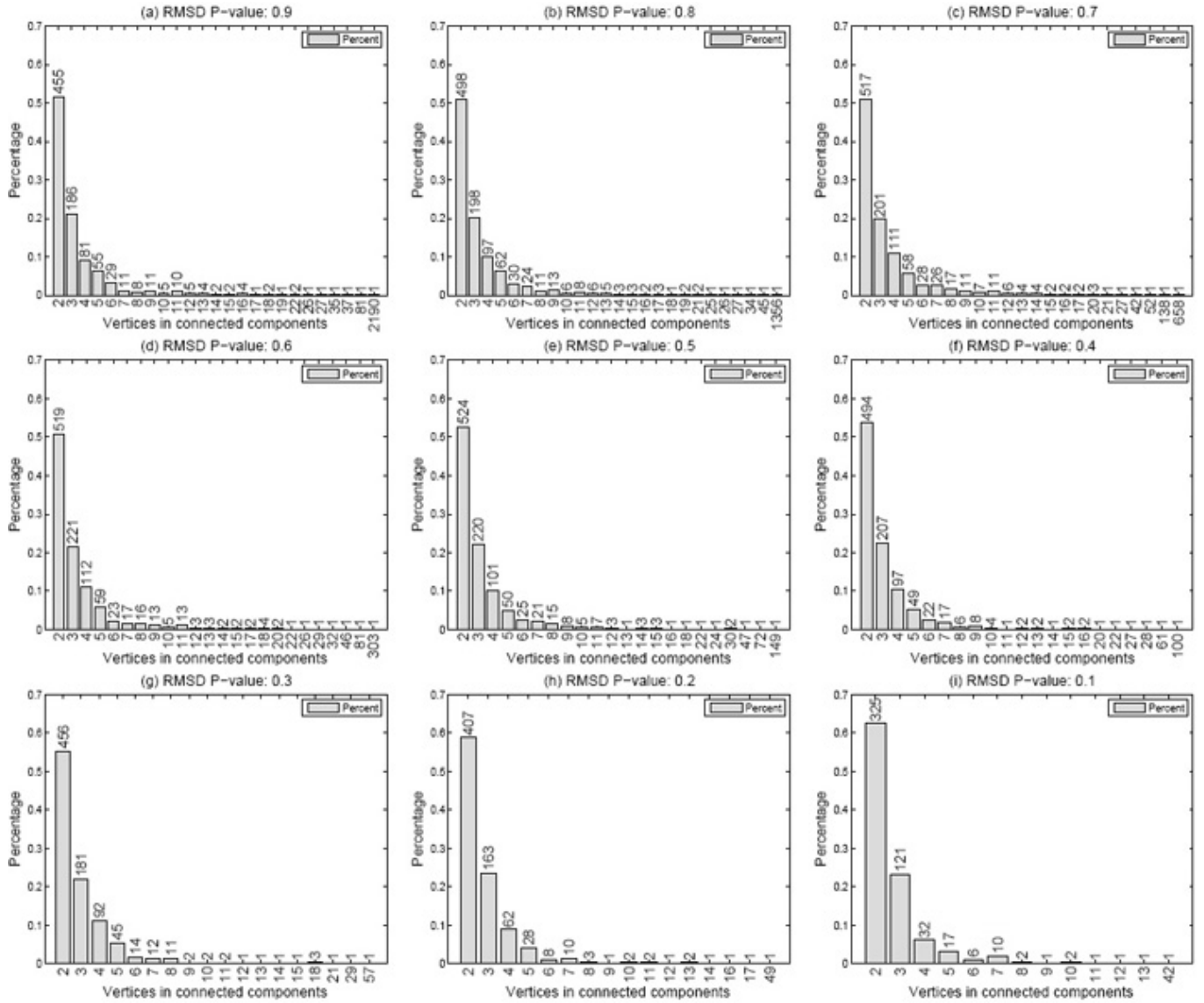

Figure 3. The percentage of non-trivial connected components with different size in the pocket similarity networks. The concrete number of the non-trivial connected components is also shown on the top of each bar individually.

B is also similar to A) are relatively satisfied well. In the network, similar pockets represent a modular architecture that they tend to be highly connected.

Secondly, the community structure in the pocket similarity network reveals the evolutionary trace of the cavity shapes for proteins from the structural perspective. Since the shapes of certain cavities would be conversed during the functional divergence, it is very important to find the conversed surface patterns. From such a viewpoint, the community structures in the network provide information to identify

the 3D structure motifs as building blocks of a protein. It should be noted that these local sequence independent motifs (pockets) are more important in finding remoter evolutionary relationships than sequences and structures. Hence, the surface patterns in the network have strong relationships to particular functional roles in organisms. It is possible to develop a function annotation or classification system for new proteins by checking their structural motifs, according to the paradigm that similar structures tend to have similar functions.

3.2. Small-World Characteristic and Scale-Free Distribution

We performed more detailed analysis in the maximum connected components by different thresholds which contain most nodes. The numbers of nodes and edges of the maximum connected components in the networks with different similarity thresholds are listed in Table **2**.

Firstly, we show that the maximum connected components have the small-world characteristic. We compared them with the random networks with equal size by two measures, i.e. the path length (L) and the clustering coefficient (C) [3]. The parameters L and C in a small-world network should satisfy two criteria: (1) L_{small_world} slightly exceeds L_{random}; (2) C_{small_world} far exceeds C_{random}. Specifically, if a node v has k_v neighbors, the maximum number of edges between these neighbors is $[k_v \times (k_v - 1)] / 2$. C_v is defined as the fraction of these possible edges that actually exist, and C is the average C_v over all nodes v. C is a measure of the local clustering within a network [6,8]. L is defined as the number of edges in the shortest path between two nodes, and averaged over all pairs of nodes. The results are listed in Table **2** respectively. It shows that all the maximum connected components are small-world. Such a tendency becomes more obvious when the network contains more nodes and edges.

Secondly, we found that the maximum connected components are scale-free networks. Scale-free networks typically have many nodes with few edges, and only a few are highly connected ones. In contrast to a random network in which the connectivity distribution obeys a Poisson distribution, the probability $P(k)$ of nodes having k edges, decays as a power law $P(k) = k^{-\gamma}$ and $log(P(k)) = -\gamma log(k)$ in a scale-free network. $P(k)$ is the distribution of the number of nodes in a given maximum connected part, k is the degree, i.e. the number of the edges per node, and γ is the degree exponent [5]. A straight fitting line on a double logarithmic is a standard way to measure the distribution and to identify γ, which is the slope of the line. We analyzed the distribution patterns for the nine maximum connected components. Fig. (**4**) plots both the connectivity distributions and the double logarithmic scale for more reliable identification of a linear fit and characteristic of a scale-free topology. We found a line to fit the double logarithmic data between the degree of number of edges per node *verse* the probability of the distribution. The regression lines of the 9 subnetworks are shown in Fig. (**5**). γ of the power-law distribution and the R^2 coefficient are also shown. From Fig. (**5**), the maximum subnetworks can be approximately characterized by power laws individually, where $P(k) = k^{-2.537}$ ($R^2 = 0.953$) and $P(k) = k^{-0.816}$ ($R^2 = 0.845$) in case of networks constructed by cRMSD p-value 0.9 and 0.1 respectively. The scaling exponent γ is gradually changed in the range from 2.537 to 0.843 with the threshold becoming much tenser. Here R^2 measures the data fitting by calculating the total variation in the data about the average. A value closer to 1 indicates a better fitting [5].

3.3. The Hub Pockets

The highly connected pockets, called hubs, are representatives of the most connected nodes in the similarity network. The particular characteristic of the nodes would provide more information for the similar concavity patterns on protein surfaces. Moreover, the physicochemical relationships among these hub pockets with the special features of the network show more valuable clues to study the similarities among surface motifs. Table **3** shows that the ten hubs in the similarity network constructed by taking the threshold as p-value 0.9.

Table 2. The Statistical Results of the Nine Maximum Connected Components in the Networks with Different Similarity Thresholds. "Node" and "Edge" are the Number of Nodes and Edges Respectively. The Other Values are the Statistical Measure of Each Maximum Connected Component

Threshold	Node	Edge	L	L_{random}	C	C_{random}
0.9	2190	2548	15.835	9.107	0.018	0.001
0.8	1356	1611	16.920	8.398	0.02	0.002
0.7	658	782	16.929	7.495	0.019	0.004
0.6	303	382	11.916	6.178	0.019	0.008
0.5	149	176	7.897	5.821	0.021	0.016
0.4	100	121	6.890	5.211	0.020	0.024
0.3	57	111	3.482	2.974	0.187	0.068
0.2	49	102	3.196	2.729	0.211	0.085
0.1	42	92	3.384	2.530	0.248	0.104

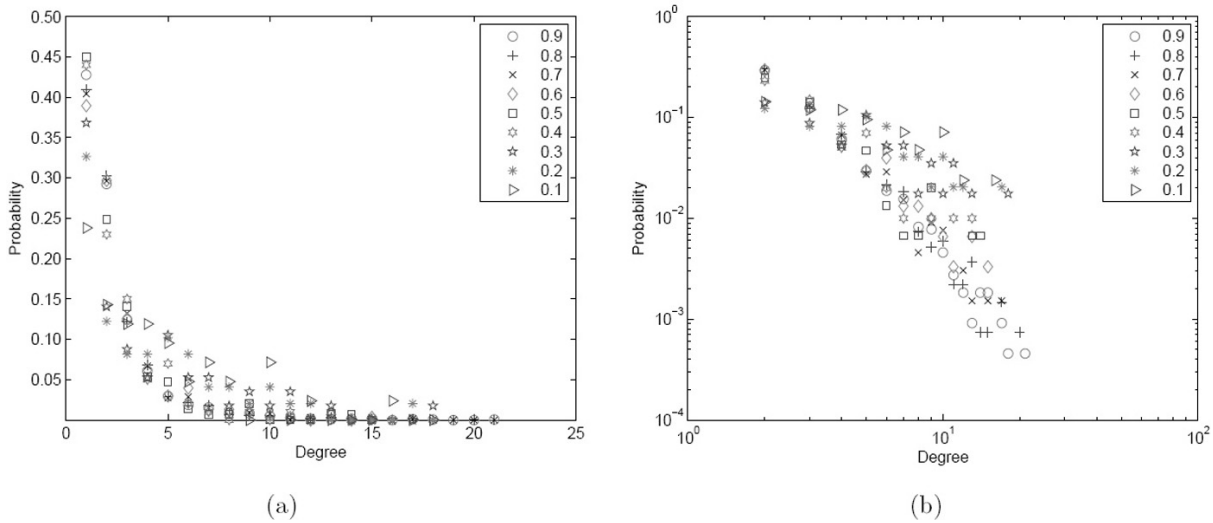

Figure 4. The degree distributions of the maximum connected components in the networks with different similarity thresholds. (a) The distributions of node connectivity $P(k)$. (b) The log-log plots.

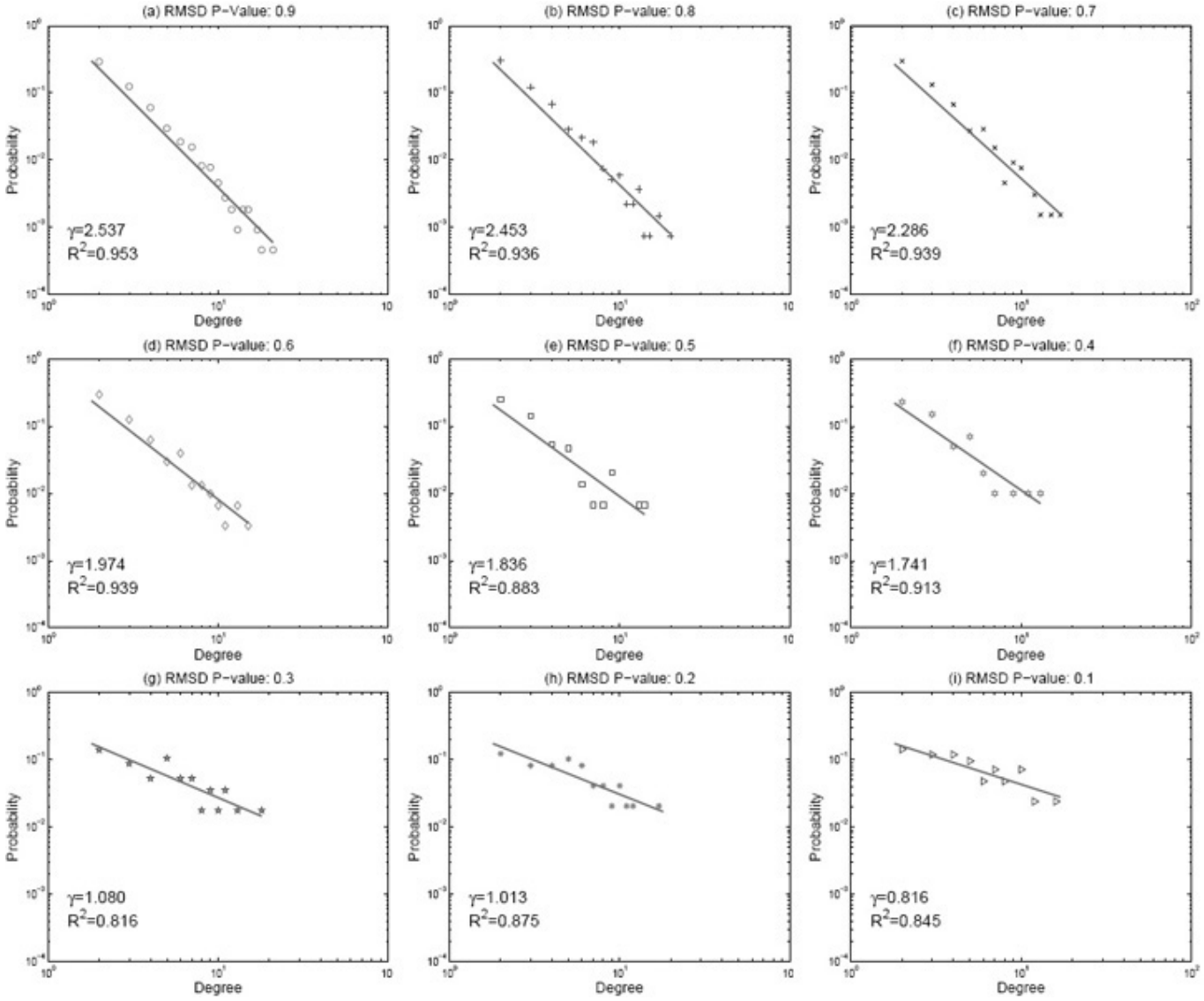

Figure 5. The regressions of the degree distribution in the maximum connected components.

Table 3. The Top 10 Most Highly Connected Hubs in the Pocket Similarity Network

Pocket	Protein	Residue Length	Volume	Degree
1dtd_43_A	metallocarboxypeptidase inhibitor	7	42.74	21
1a97_76_B	xanthine-guanine phospho-ribosyl transferase	39	1443.01	18
1exn_58_B	5'-exonuclease (Se-Met labelled protein)	38	1547.7	17
2jdx_52_A	glycine amidinotransferase, deletionmutant atdeltam302	33	708.99	17
1byi_19_0	dethiobiotin synthase	15	355.77	15
2ebo_28_A	ebola virus envelope glycoprotein	14	210.61	15
1im0_37_A	outer membrane phsopholipase	65	2067.34	15
1k8u_12_A	calcium-free (or apo) human s100a6	19	787.1	15
1a4i_59_B	methylenetetrahydrofolate dehydrogense	14	308.84	14
1aym_123_A	human rhinovirus 16 coat protein	25	570.08	14

In Table **3**, the residue length of a pocket is the number of amino acid residues concatenated from non-sequential positions composing pocket on the primary chain of the protein. The volume units are $Å^3$. From Table **3**, we can find that the residue lengths and the volumes of these hub pockets do not have significant correlation relationships. It means that these hubs in the pocket similarity network distribute diversely and do not have similarity in size or volume. The hubs are similar to relatively more pockets in the network, and tend to work as the "center" of some communities. Then the diversity among hubs indicates that the community structures in the similarity network are diverse. This provides more evidences that the similar pockets tend to gather together and naturally constitute community structures in the network. The degree of the top ten hubs demonstrate that it has no determinate implication that the pockets with smaller sizes can find similar pockets more easily in the pocket database. Moreover, the hub proteins are from different species, and they easily interact with other protein domains to perform functions. This also implies that the hub proteins and the surface motifs of hub pockets are conserved during evolution and they are really the representatives for protein surface patterns. In addition, we can conclude that the similarity among the pockets constitute a kind of special complex system with the features of the hub pockets. Further analysis on the functional features of components in the pocket similarity network is required to elucidate their biological implications.

4. DISCUSSION AND CONCLUSION

It is well known that proteins interact with other molecules to perform their biological functions. The key factors in all these interactions are the shapes and chemical properties on protein's surfaces. The surface is generally irregular, containing many pockets, which show high relevance to binding sites. In this work, the structural similarity among the protein surface patterns has been explored by topological properties of the pocket similarity network. The comprehensive analysis of the pocket similarity network shows that the similarity among the protein surface patterns has the community structure, which provides evidence that the protein surface pattern is conserved in evolution and the similar pockets in a cluster may have similar biological functions. We also found the maximum connected components in the pocket similarity networks are small-world and scale-free as other complex networks of protein universe [5,9]. The results give a direct answer to the problem about the features of the similarity network proposed in [35] and provide more detailed information on the whole network's architecture because the maximum subnetwork relatively contains most components of the whole network.

The structural similarity among these surface patterns also provides valuable information to detect the conserved structural features and further functions. In this paper we main examined the aspect of surface structures of proteins. In the future, however, we will concentrate on the functional features of the similarity network. The analysis of the physicochemical feature of the hub pockets implies that we can analyze and demonstrate much more functional implications from the similarity network model, which provides new insights into structural genomics and have great potential for applications on functional genomics. The community structure of the similarity network can then be explored to develop a straightforward classification method for dividing the similar pockets into small groups. Furthermore, these compositional and evolutionary information at conserved structural motifs can be compiled into a library of functional templates after these pocket groups are functionally annotated [15,37]. The method to cluster the pockets based on the concept of pocket similarity network and the functional classification of these pockets are in progress and will be presented in another paper.

One of the grand challenges in systems biology is to build a complete and high-resolution description of molecular topography and connect molecular interactions with physiological responses. By studying the relationships between protein pockets instead of individual pockets, we aim to reveal the holistic protein structural relationships and integrate all the existing 3D structural data in PDB. Such work

will provide the groundwork, which may lead to further research to understand the complex protein world.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 10631070 and No. 60503004, the JSPS and NSFC under JSPS-NSFC collaboration project, and the Chinese Ministry of Science and Technology under Grant No. 2006CB503905. Part of the authors is also supported by the Knowledge Innovation Program of the Chinese Academy of Sciences. The authors are grateful to the anonymous referees for their valuable comments and suggestions in improving the presentation of the earlier version of the paper. The authors wish to thank Dr. Jie Liang for providing the data of CASTp and pvSOAR databases.

ABBREVIATIONS

3D = Three dimensional

PDB = Protein data bank

CASTp = Computed atlas of surface topography of proteins

pvSOAR = Pocket and void surfaces of amino acid residues

cRMSD = Coordinate root mean square distance

GO = Gene ontology

GOA = Gene ontology annotation

REFERENCES

[1] Chen L., Wu L.Y., Wang Y. and Zhang X.S. (**2006**) *Proteins, 62,* 833-837.
[2] Wang Y., Joshi T., Zhang X.S., Xu D. and Chen L. (**2006**) *Bioinformatics, 22,* 2413-2420.
[3] Watts D.J. and Strogatz S.H. (**1998**) *Nature, 393,* 440-442.
[4] Barabasi A.L. and Albert R. (**1999**) *Science, 286,* 509-512.
[5] Greene L.H. and Higman V.A. (**2003**) *J. Mol. Biol., 334,* 781-791.
[6] Strogatz S.H. (**2001**) *Nature, 410,* 268-276.
[7] Girvan M. and Newman M.E. (**2002**) *Proc. Natl. Acad. Sci. USA, 99,* 7821-7826.
[8] Wuchty S. (**2001**) *Mol. Biol. Evol., 18,* 1694-1702.
[9] Amitai G., Shemesh A., Sitbon E., Shklar M., Netanely D., Venger I. and Pietrokovski S. (**2004**) *J. Mol. Biol., 344,* 1135-1146.
[10] Rao F. and Caflisch A. (**2004**) *J. Mol. Biol., 342,* 299-306.
[11] Eisenberg D., Marcotte E.M., Xenarios I. and Yeates, T.O. (**2000**) *Nature, 405,* 823-826.
[12] Orengo C.A., Todd A.E. and Thornton J.M. (**1999**) *Curr. Opin. Struct. Biol., 9,* 374-382.
[13] Lewis M. and Rees D.C. (**1985**) *Science, 230,* 1163-1165.
[14] Schmitt S., Kuhn D. and Klebe G. (**2002**) *J. Mol. Biol., 323,* 387-406.
[15] Chen L., Wu L.Y., Wang Y., Zhang S. and Zhang X.S. (**2006**) *BMC Structural Biology, 6,* 18.
[16] Hendlich M., Rippmann F. and Barnickel G. (**1997**) *J. Mol. Graph Model,15,* 359-63,389.
[17] Kleywegt G.J. and Jones T.A. (**1994**) *Acta Crystallogr D. Biol. Crystallogr, 50,* 178-185.
[18] Liang J., Edelsbrunner H. and Woodward C. (**1998**) *Protein Sci., 7,* 1884-1897.
[19] Laskowski R.A., Luscombe N.M., Swindells M.B. and Thornton J.M. (**1996**) *Protein Sci., 5,* 2438-2452.
[20] Binkowski T.A., Adamian L. and Liang J. (**2003**) *J. Mol. Biol., 332,* 505-526.
[21] Ferre F., Ausiello G., Zanzoni A. and Helmer-Citterich M. (**2005**) *BMC Bioinformatics, 6,* 194.
[22] Nayal M. and Honig B. (**2006**) *Proteins, 63,* 892-906.
[23] Laurie A.T. and Jackson R.M. (**2005**) *Bioinformatics, 21,* 1908-1916.
[24] Jones S. and Thornton J.M. (**1996**) *Proc. Natl. Acad. Sci. USA, 93,* 13-20.
[25] Jones S. and Thornton J.M. (**1997**) *J. Mol. Biol., 272,* 121-132.
[26] Jones S. and Thornton J.M. (**1997**) *J. Mol. Biol., 272,* 133-143.
[27] Stark A., Sunyaev S. and Russell R. (**2003**) *J. Mol. Biol., 326,* 1307-1316.
[28] Binkowski T.A., Joachimiak A. and Liang J. (**2005**) *Protein Sci., 14,* 2972-2981.
[29] Tseng Y.Y. and Liang J. (**2006**) *Mol. Biol. Evol., 23,* 421-436.
[30] Kuhn D., Weskamp N., Schmitt S., Hullermeier E. and Klebe G. (**2006**) *J. Mol. Biol., 359,* 1023-1044.
[31] Binkowski T.A., Naghibzadeh S. and Liang J. (**2003**) *Nucleic Acids Res., 31,* 3352-3355.
[32] Binkowski T.A., Freeman P. and Liang J. (**2004**) *Nucleic Acids Res., 32,* W555-W558.
[33] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (**2000**) *Nucleic Acids Res., 28,* 235-242.
[34] Hobohm U. and Sander C. (**1992**) *Protein Sci., 1,* 409-417.
[35] Zhang Z. and Grigorov M.G. (**2006**) *Proteins, 62,* 470-478.
[36] Newman M.E. (**2004**) *Phys. Rev. E., 69,* 066133.
[37] The Gene Ontology Consortium. (**2000**) *Nature Genet., 25,* 25-29.

Received: ?????????????? Revised: ?????????????? Accepted: ??????????????????