The First International Symposium on Optimization and Systems Biology (OSB'07)
Beijing, China, August 8–10, 2007
Copyright © 2007 ORSC & APORC，pp. 204–212

An Approach for Clustering Protein Pockets into Similar Groups*

Zhi-Ping Liu[1,2,†] Ling-Yun Wu[1,‡] Yong Wang[1]
Xiang-Sun Zhang[1] Luonan Chen[3,4]

[1] Institute of Applied Mathematics, 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] Department of Electronics, Information and Communication Engineering
Osaka Sangyo University, Osaka 574-8530, Japan
[4] Institute of Systems Biology, Shanghai University, Shanghai 200444, China

Abstract In this work, we propose a network-based approach to cluster the protein pockets into similar groups in a database level. A pocket similarity network is constructed to describe structural similarity relationships among the pockets from the systematic perspective, which possesses the community structures and can can be utilized to develop a direct method to cluster the pockets by partitioning the network to small communities, which correspond to pocket groups individually. As a first step, we join the pockets into structurally similar pocket groups via a hierarchical process guided by maximizing a widely used modularity measurement Q. Then we analyze the functional similarity underlying every divided pocket groups. As a result most of the pockets in the same group are identified to share similar functions. These results show that our clustering method is effective and efficient to reveal biologically meaningful pocket groups regard to functional consistence.

Keywords Protein pockets; clustering; pockets similarity network; functional genomics; systems biology.

1 Introduction

It is well known that protein functions are mainly determined by its physical, biochemical and geometric properties of structural surface [1]. These surface regions, e.g. pockets or clefts, provide specialized environments for biological activity, thus their underlying three-dimensional shapes and physicochemical textures are closely related to protein functions [2, 3]. Grouping the structurally similar surface regions is useful to extract functionally conserved spatial patterns during evolution. It can also provide important insights into the biochemical relationships between functions

[*] This work was supported by National Natural Science Foundation of China (NSFC) under Grant No. 10631070 and No. 60503004. Part of the authors are also supported by the 973 Program (Grant No. 2006CB503910) from the Ministry of Science and Technology of China.

[†] Email: zpliu@amss.ac.cn.

[‡] Corresponding author. Email: lywu@amt.ac.cn.

and structural motifs, in particular based on the assumption: the similar structural features imply similar functions.

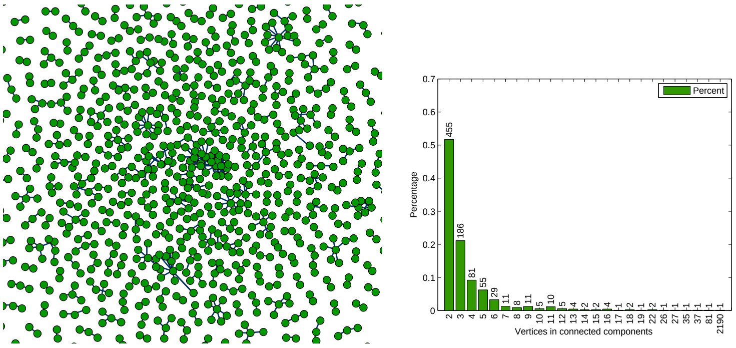
A straightforward way to perform clustering is to introduce the concept of network. Analyzing and using network properties can characterize both the whole system and its individual components [4], hence such a strategy has been widely applied in many disciplines. In particular, the network analysis has attracted much attention on the area of systems biology due to wide availability of high-throughput data, such as the protein-protein interactions, the interactions among families of protein domains, and the amino acid contacts within protein structures [5, 6, 7, 8, 9, 10]. One important feature of these networks is their community structure, which is viewed as the gathering of vertices in groups, within which the network connections are dense, but between which the links are sparse. The community structure often relates to valuable components of the network [11].

In this work, we aim to develop a direct classification procedure for clustering the pockets into small groups based on a similarity network, which is introduced to systematically describe the similarities among the pockets [12]. We found that the pocket similarity network possesses the feature of community structure. The architecture of the similarity network implicates that the feature can be directly utilized as a criterion in the clustering approach. After briefly reviewing the topological features of the similarity network, we split the pockets in a recursion manner into small components. Then the quality of clustering is assessed by an extensively used modularity measurement and the functional relationships among the pockets in every detected group. The experimental results provide an evidence that the proposed method is effective to cluster the pockets, and the pocket groups are biologically meaningful. Furthermore, our idea of the network model and the network partitioning method can be easily extended to clefts or other protein structural motifs in bioinformatics.

2 Community structure of the pocket similarity network

Recently, we introduced the similarity network model to systematically describe the structural similarity among protein pockets [12], and to detect the features of the network comprehensively. Specifically, we use the proteins in PDB_SELECT25 to remove the redundancy in PDB, in which the proteins have low sequence similarity (less than 25%). This indicates that they come from different protein families. We collect all the pockets of proteins in PDB_SELECT25 from CASTp [13] database (78925 pockets). 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 derived from pvSOAR [14] database. When querying one pocket in pvSOAR, it would hit some similar pockets satisfying the given threshold. The pvSOAR database compares the pockets in CASTp [14] in an all-against-all way. We use a threshold, structural cRMSD (coordinate root mean square distance) p-value 0.9, to choose the connections. As a result an edge in pocket similarity network links two structurally similar pockets. The isolated nodes in the network are discarded, because they can not be used to find similar pockets in the whole pocket library. Figure

1 gives an example of the constructed pocket similarity network. We found that the similarity network possesses the community structure feature, i.e. the similar pockets tend to cluster together and constitute many communities in the sparse network spontaneously (As shown in Figure 1). We also analyzed the other features such as the small-world behavior and scale-free property underlying the network. The readers can refer to [12] for the detailed analysis of the network properties. In the present work, we utilize the network features to cluster the pockets into small groups.



(a) (b)

Figure 1: (a) A part of the pocket similarity network (b) The percentage of connected components with different size in the pocket similarity network. The concrete number of the connected components is also shown on the top of each bar individually.

3 Clustering the pockets into small groups

The community structure of the pocket similarity network provides us a straightforward and simple way to cluster the pockets into small groups. Figure 2 shows the framework and illustration of the process to partition the network. In fact it is a recursion algorithm using a community structure detecting algorithm as a subprocess.

Our method to partition the network is based on a widely used concept of modularity measurement Q, which is a quality function to measure whether a particular division is meaningful. Q is defined as $Q = \sum_i (e_{ii} - a_i^2)$, where e_{ij} is the fraction of edges in the network that connect vertices in community i to those in community j, and $a_i = \sum_j e_{ij}$. Then Q is the fraction of edges that fall within communities, minus the expected value of the same quantity of edges falling at random without regard to the community structure. If a particular division gives no more within-community edges than which would be expected by random chance, the value of Q would be 0. The value of Q approaches 1, which is the maximum value, more closely indicates more strong community structure. Generally, Q values for networks typically fall in the range from about 0.3 to 0.7 [15]. Detecting the partition of groups that maximizes Q is believed to be a *NP-hard* problem, which makes a brute force exploration impossible for large scale networks having dozens of vertices. However, there is a fast algorithm to create the hierarchy in an agglomerative strategy to maximize the Q

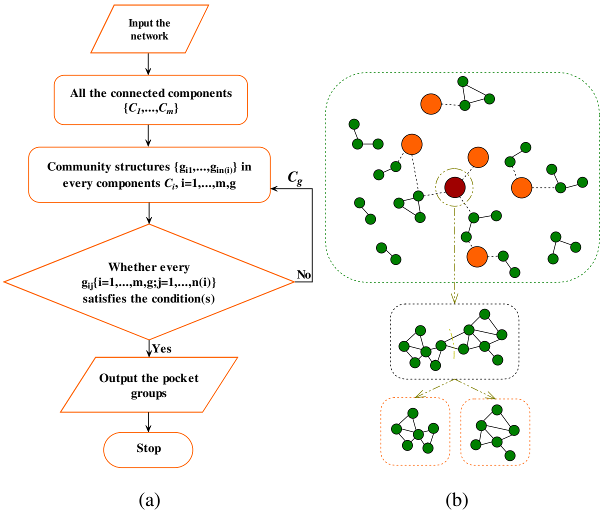

Input the network
All the connected components $\{C_1,\ldots,C_m\}$
Community structures $\{g_{i1},\ldots,g_{in(i)}\}$ in every components C_i, i=1,...,m,g
C_g
Whether every $g_{ij}\{i=1,\ldots,m,g;j=1,\ldots,n(i)\}$ satisfies the condition(s)
No
Yes
Output the pocket groups
Stop
(a)
(b)

Figure 2: The hierarchical clustering process is designed to divide the network into small communities. (a) The flow chart of the whole algorithm. (b) An illustration graph of the recursion process. The big nodes mean the communities that can be divided in a recursive way.

[15]. At first, the algorithm regards every single node as a cluster, then two clusters are merged into one to maximize the the increment of modularity. The process is repeated until only one cluster remains. Clearly, for a network of n vertices, there would be $n-1$ steps for such joining. The algorithm is very efficient and widely used [16].

The main idea is to partition the network into small clusters, which correspond to the pocket groups individually. If a community does not satisfy the stop criteria, for example, the community is still rather large (contains many nodes) and/or possesses the obvious community structure underlying the cluster, we continue to divide it to smaller components. We modify the fast algorithm of detecting the community structure in a large complex network [15] to a subprocess to partition the similarity network to small clusters. Firstly, we find out all the connected components in the whole network. Then we use the fast algorithm to detect community structure in each connected component respectively. In the connected component C_i, the subprocess is used to find out the communities that maximize the Q_i of the subnetwork. Finally, for each community, if its size is below a given value and/or the modularity Q is below a given threshold, the algorithm is stopped and the community is regarded as a pocket group. Otherwise, we continue to partition the community into smaller communities.

The value of the modularity Q is use to measure our divisions. During the whole process of dividing the network, we record the changing of Q. The bigger modularity Q is, the more obviously we can divide the network into smaller communities. Thus we choose the detecting value of Q as the stop criterion of our algorithm. Moreover, we also analyze the functional consistence in every group. The high modularity Q of the clusters and the functional features underlying the groups show that the divided groups are biologically meaningful.

4 The clusters of pocket group

The pocket similarity network have 5387 vertices and 4943 edges. From the statistics of the similarity network shown in Figure 1, the network contains 880 connected components. Most of the components contain a few nodes. The maximum connected component of the similarity network contains 2190 vertices with 2548 edges. The second largest connected component contains 81 vertices with 83 edges. Figure 1 also shows that the similar pockets are naturally clustered together. Noting that small connected components may contain less information, in our numerical experiments, we take 81 as the threshold of the size of pocket groups in the partition of the two largest connected components.

Table 1: The number of the clusters of pockets when we choose 81 as the size threshold for every group.

Connected component	Vertex	Edge	Num of clusters	Max size	Min size	Mean size	Max Q
Largest	2190	2548	49	117	19	44.694	0.935
Second largest	81	83	8	13	4	10.125	0.776
The rest (self-clustered)	3116	2312	878	37	2	3.549	–

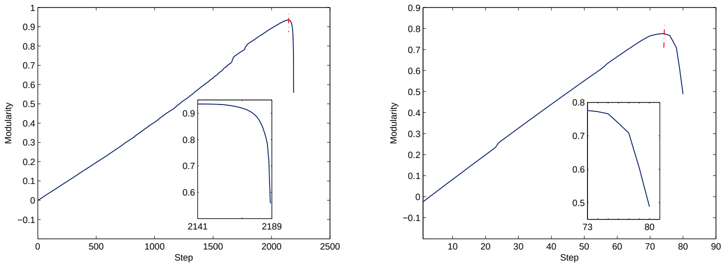
Table 1 records the statistics of the components after the first level clustering procedure. In Table 1, the largest connected component of the network is partitioned into 49 small communities after 2141 joining steps. The second largest connected component is divided into 8 clusters after 73 steps. The modularity Q measures the significance of the community structure of the partitioned network. Figure 3 records the change of Q. The cut-off point of the joining steps with the maximum modularity is also shown in Figure 3. Subfigures (a) and (b) correspond to the largest and the second largest connected components respectively.

In the first level clusters of the largest connected component, there are two clusters whose sizes are bigger than the given threshold, 81. We continue to run the clustering procedure on the two clusters and the results are shown in Table 2.

Table 2: The smaller clusters by further partitioning the two first level clusters in the largest connected component.

Cluster	Vertex	Edge	Num of clusters	Max size	Min size	Mean size	Max Q
1	117	147	12	28	3	9.75	0.670
2	95	165	10	18	3	9.5	0.503

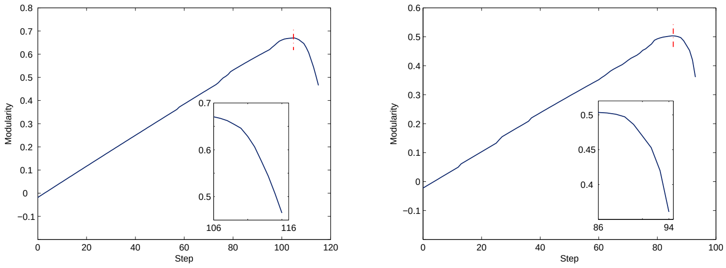
The maximum cluster in the first level communities after dividing the largest connected component contains 117 nodes and 147 edges. We cut the hierarchy tree of the partition when the modularity Q reaches the maximum 0.670. The cluster is divided into 12 smaller clusters. Figure 4 (a) records the change of Q and the cut-off



(a) (b)

Figure 3: The changing modularity Q by joining of vertices to clusters. (a) The largest connected component. (b) The second largest connected component.

point on the curve. In the similar way, the second largest cluster is divided into 10 smaller communities. Figure 4 (b) shows the results.



(a) (b)

Figure 4: The changing modularity Q by joining vertices to clusters. (a) Changing Q of partitioning the largest cluster. (b) The results of the second largest cluster.

From the above two tables, we partition the pocket similarity network to $(878+47+8+12+10)=955$ clusters, which are regarded as pocket groups. Of course, we can change the threshold of the maximal size of the clusters or combine the threshold with the Q value of the potential division. The number of pocket groups will increase if a smaller value is chosen. For instance, if we choose 37 as the threshold of the maximal size of clusters, the similarity network would be divided into 1137 small communities.

5 Functional similarity lies in the pocket groups

Based on the assumption that similar structures imply similar functions for proteins, we investigate the functional similarity in these pocket groups by annotating the GO functions of the protein in which the pocket are located. In the 955 pocket groups, there are 816 (85.45%) groups in which at least two pockets have GO terms, i.e. there are more than two proteins containing the pockets in the groups have GO annotations in GOA database [17]. We call the part of pockets with GO annotations in every group as the GOA part. In the 816 pocket groups, there are 99 (12.13%) groups which have at least one same GO term in their GOA part. There are also 191 (23.41%) pocket groups containing significant common (2/3) GO terms in their GOA part, i.e. 2/3 of the annotated pockets of the GOA part in every pocket groups have common GO terms. And there are 578 (70.83%) pocket groups with significant common (1/2) GO terms. Table 3 shows the functional similarity among the pocket groups of the size from 2 to 6 (733 (89.83%) of the 816 groups, and the full list is available upon request). The functional similarity among every pocket groups provides more evidences: the similar pockets have similar functions, and the pocket groups are functionally important structural motifs for proteins. Conversely, the results also prove that the clustering method based on the community structure is efficient to group the pockets.

Table 3: The functional similarity among the pocket groups. We annotate the pockets by the GO terms of the proteins containing them. '-' indicates the value that we need not calculate.

Group size	Number	GOA status	Common GO terms					Percentage
		(size of GOA part : number of groups)	6	5	4	3	2	
2	455	2: 334	-	-	-	-	56	16.77%
3	188	3: 130	-	-	-	16	52	44.25%
		2: 44	-	-	-	-	9	
4	84	4: 53	-	-	5	11	23	73.17%
		3: 20	-	-	-	1	9	
		2: 9	-	-	-	-	3	
5	57	5: 34	-	4	2	9	16	66.67%
		4: 16	-	-	1	2	23	
		3: 4	-	-	-	0	3	
		2: 1	-	-	-	-	1	
6	33	6: 18	0	2	3	2	9	65.63%
		5: 8	-	0	0	0	0	
		4: 6	-	-	0	1	4	
		3: 0	-	-	-	-	-	
		2: 0	-	-	-	-	-	

6 Discussion and Conclusion

Protein's functions are carried out through interacting and binding other molecules on its surface region. The surface always contains many pockets which have shown high relevance to active sites. The classification of these patterns would provide valuable insights to the relationship between protein surface and function. In this work, we proposed a novel method to cluster the pockets to small groups based on community structure feature of the similarity network directly. We also provided measurements to the clustering scheme in terms of the modularity Q and revealed the implications of functional similarity among the pockets in the same group, which provides evidences that these pocket groups are the clusters with both structural and functional similarity.

The structural similarity features among the pockets in a database level have been explored by topological properties of the pocket similarity network [12]. Our clustering method is based on the attribution of the similarity among the pockets. This is an entirely new clustering scheme which stresses importance on the structural similarity among the pockets, although there are some other clustering methods, such as K-means [18]. Directly using the traditional clustering methods may have risk in the structural data of the pockets. When we calculate the means of RMSD difference of several pockets, the risk is that we may lose the essential implications of the value of structural difference. In the friable case, we just used the similarity relationship among the pockets and used the community detecting algorithm, we can enucleate the similar pocket groups. The method can easily extend to other objects in protein's universe. A direct comparison and measurement among these clustering methods is a challenging task for us in the future. Moreover, these pocket groups would have potentially important applications. One direction is to develop a library of structural motifs using these pocket groups. When the concrete functions of a group are identified, the group is a functional template and might be used to function prediction, drug design and other bioengineering. The physicochemical features of the pockets are important for understanding the functional sites, and the evolutionary information can also be derived from the multiple pocket sequence alignment, which are our undergoing work.

In conclusion, we developed a novel clustering scheme to assign the pockets into small groups in a database level. The method is based on the unique features of the similarity network, which maps the structural relationships in a systematic way. We modified the community structure detecting algorithm to partition the network to small clusters. The high modularity Q of the division provides evidence that the partition considers the topology information of the similarity network efficiently. And the functional similarity within the pocket groups shows that the groups are biologically meaningful. The presented method can be extended to other problems or definitions in structural systems biology, and the simulation results demonstrated that the functionally important pocket groups can have important applications in functional genomics.

Acknowledgments

Thanks due to Dr. Jie Liang for providing us the data of CASTp and pvSOAR databases. We also thank Dr. Mark Newman for providing us the software of detecting the community.

References

[1] Schmitt S., Kuhn D., Klebe G.: A new method to detect related function among proteins independent of sequence and fold homology. J. Mol. Biol., **323** (2002) 387–406.

[2] Laskowski R.A., Luscombe N.M., Swindells M.B., Thornton J.M.: Protein clefts in molecular recognition and function. Protein Sci., **5** (1996) 2438–2452.

[3] Binkowski T.A., Adamian L., Liang J.: Inferring functional relationships of proteins from local sequence and spatial surface patterns. J. Mol. Biol., **332** (2003) 505–526.

[4] Strogatz S.H.: Exploring complex networks. Nature, **410** (2001) 268–276.

[5] Wuchty S.: Scale-free behavior in protein domain networks. Mol. Biol. Evol., **18** (2001) 1694–1702.

[6] Greene L.H., Higman V.A.: Uncovering Network Systems Within Protein Structures. J. Mol. Biol., **334** (2003) 781–791.

[7] Rao F., Caflisch A.: The protein folding network. J. Mol. Biol., **342** (2004) 299–306.

[8] Chen L., Wu L.Y., Wang Y., Zhang X.S.: Inferring protein interactions from experimental data by association probabilistic method. Proteins, **62** (2006) 833–837.

[9] Chen L., Wu L.Y., Wang Y., Zhang S., Zhang X.S.: Revealing divergent evolution, identifying circular permutations and detecting active-sites by protein structure comparison. BMC Structural Biology, **6** (2006) 18.

[10] Wang Y., Joshi T., Zhang X.S., Xu D., Chen L.: Inferring gene regulatory networks from multiple microarray datasets. Bioinformatics, **22** (2006) 2413–2420.

[11] Girvan M., Newman M.E.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA, **99** (2002) 7821–7826.

[12] Liu Z.P., Wu L.Y., Wang Y., Zhang X.S., Chen L.: Analysis of protein surface patterns by pocket similarity network. Protein & Peptide Letters, (2007) in press.

[13] Binkowski T.A., Naghibzadeh S., Liang J.: CASTp: Computed Atlas of Surface Topography of proteins. Nucleic Acids Res., **31** (2003) 3352–3355.

[14] Binkowski T.A., Freeman P., Liang J.: pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. Nucleic Acids Res., **32** (2004) W555–W558.

[15] Newman M.E.: Fast algorithm for detecting community structure in networks. Phys. Rev. E., **69** (2004) 066133.

[16] Clauset A., Newman M.E., Moore C.: Finding community structure in very large networks. Phys. Rev. E., **70** (2004) 066111.

[17] The Gene Ontology Consortium.: Gene Ontology: tool for the unification of biology. Nature Genet., **25** (2000) 25–29.

[18] Jain A.K., Murty M.N., Flynn P.J.: Data clustering: a review. ACM Computing Surveys, **31** (1999) 264–323.