Protein cavity clustering based on community structure of pocket similarity network

Zhi-Ping Liu

Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing 100080, China

Graduate University of Chinese Academy of Sciences,
Beijing 100049, China
E-mail: zpliu@amss.ac.cn

Ling-Yun Wu*, Yong Wang
and Xiang-Sun Zhang

Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing 100080, China
E-mail: lywu@amt.ac.cn
E-mail: ywang@amss.ac.cn
E-mail: zxs@amt.ac.cn
*Corresponding author

Luonan Chen

Department of Electronics,
Information and Communication Engineering,
Osaka Sangyo University, Osaka 574-8530, Japan

Institute of Systems Biology,
Shanghai University, Shanghai 200044, China
E-mail: chen@eic.osaka-sandai.ac.jp

Abstract: Functions of a protein are mainly determined by its structure. Surface cavities, also called pockets or clefts, are ordinarily regarded as potentially active sites where the protein carries out the functions. Clustering these pockets is a challenging task in structural genomics. In this paper, we introduce pocket similarity network which possesses the feature of community structure to systematically describe structural similarity among pockets, then a straightforward classification scheme is developed based on this special feature. The surface pockets are clustered into structurally similar pocket groups via a hierarchical process. We identify these small pocket groups as structural templates which represent similar functions in diverse proteins. The experimental results show that our clustering method is effective, and the identified pocket groups are biologically meaningful in terms of their functional features.

Keywords: protein surface pattern; clustering; pocket similarity network; structural motif; functional genomics.

Reference to this paper should be made as follows: Liu, Z-P., Wu, L-Y., Wang, Y., Zhang, X-S. and Chen, L. (2008) 'Protein cavity clustering based on community structure of pocket similarity network', *Int. J. Bioinformatics Research and Applications*, Vol.

Biographical notes: Zhi-Ping Liu is a PhD candidate in Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China. His current working interests include bioinformatics, systems biology and combinatorial optimisation.

Ling-Yun Wu obtained his PhD Degree in Operations Research in 2002. Currently, he is an Associate Professor of Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences. He is also a research fellow of Center of Bioinformatics in AMSS. His expertise includes bioinformatics and algorithm design and analysis.

Yong Wang obtained his PhD Degree in Operations Research in 2005. Currently, he is an Assistant Professor of Institute of Applied Mathematics, Academy of Mathematics and Systems Science, CAS. He is also a research fellow in Center of Bioinformatics in AMSS. His interests are bioinformatics and systems biology.

Xiang-Sun Zhang is a Full Professor of Academy of Mathematics and Systems Science, CAS. He is the Director of Center of Bioinformatics in AMSS. His interests are bioinformatics, systems biology and mathematical programming. Now, he is also the Honorary President of Operations Research Society of China.

Luonan Chen received his PhD Degree in 1991 from Tohoku University, Sendi, Japan. Currently, he is a Professor in the Department of Electronics, Information and Communication Engineering, Osaka Sangyo University, Japan. He is also the Director of Institute of Systems Biology in Shanghai University, China. His research interests include systems biology and dynamic systems. He is an IEEE Senior Member.

1 Introduction

Generally a protein performs its biological functions by interacting with other molecules. It is well known that functions of a protein are mainly determined by its physical, biochemical and geometric properties of structural surface (Liang et al., 1998; Schmitt et al., 2002; Ferre et al., 2004). These surface regions, e.g., pockets or clefts, provide specialised environments for biological activity, thus their underlying three-dimensional shapes and physicochemical textures are closely related to protein functions (Laskowski et al., 1996; Jones and Thornton, 1997; Binkowski et al., 2003; Stark et al., 2003; Binkowski et al., 2005; Kinoshita and Nakamura, 2005; Laurie and Jackson, 2005). 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 and structural motifs, in particular based on the assumption that the similar structural features imply similar functions.

To group the protein surface patterns naturally by their structural similarity, one possible way is to introduce the concept of networks, which can easily describe the complicated relationships, by exploiting well-known research results in the area of complex networks and graph theory. Analysing and using network properties can characterise both the whole system and its individual components (Watts and Strogatz, 1998; Barabasi and Albert, 1999; Strogatz, 2001), hence such a strategy has been widely applied in many disciplines. In particular, the network analysis is a basic tool and has attracted much attention in the area of systems biology to deal with the wide availability of high-throughput data, such as the protein-protein interactions, the interactions among protein domain families, and the amino acid contacts within protein structures (Wuchty, 2001; Greene and Higman, 2003; Rao and Caflisch, 2004). One important feature of the network is its community structure. The community structure is viewed as the gathering of nodes 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 (Newman, 2004; Zhang et al., 2007).

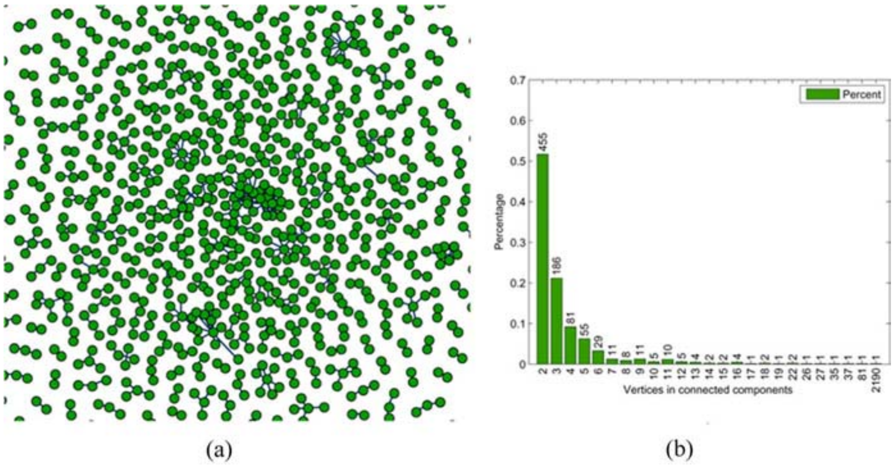
In this paper, we intend to develop a simple and effective classification procedure for clustering the protein pockets into small groups based on a similarity network which was introduced to systematically describe the similarities among the protein pockets (Liu et al., 2008). In the previous work, we found that the pocket similarity network possesses a unique feature of community structure. The architecture of the similarity network implicates that the feature can be directly utilised as a criterion in the clustering approach. After briefly reviewing the topological features of the similarity network, we describe the procedure to cluster the pockets into small groups. Then the quality of clustering is assessed by an extensively used measurement and the functional relationships among the pockets in every detected group. The statistics results show that the proposed method is effective to cluster the pockets, and the pocket groups are biologically meaningful. Furthermore, the idea of the network modelling and the network partitioning method can be easily extended to clefts or other protein structural patterns in bioinformatics.

2 Methods

2.1 The pocket similarity network and its community structure

Recently, we introduced the similarity network model to systematically describe the structural similarity relationships among protein pockets, and analysed the properties of the network comprehensively (Liu et al., 2008). The proteins in PDB_SELECT 25, in which chains have low sequence similarity (less than 25%), are used in the experiments in order to eliminate the most homologous redundancy in PDB (Hobohm and Sander, 1992; Berman et al., 2000). All the pockets of proteins in PDB_SELECT 25 are collected from CASTp database (Binkowski et al., 2003) and a pocket similarity network is constructed. In the network, each pocket is represented by a node, and 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 database (Binkowski et al., 2004). When querying one pocket in pvSOAR, it would hit some similar pockets satisfying the given threshold since the pvSOAR database contains the all-against-all similarity scores of the pockets in CASTp (Binkowski et al., 2004). We use a threshold provided by pvSOAR, the structural cRMSD (coordinate root mean square distance) p-value 0.9, to choose the connections in the similarity network. Namely, an edge in pocket similarity network links two structurally similar pockets with cRMSD p-value smaller than 0.9. The isolated nodes in the network are useless and discarded. Figure 1 shows part of the network. We found that the similarity network possesses the community structure feature, i.e., the similar pockets tend to cluster together and constitute the communities in the network spontaneously. The mechanism of the community structure feature of the similarity network lies in the special similarity metric among the pockets. The features provide implications that the surface motifs are conserved during evolution. We also analysed the other features such as the small-world behaviour and scale-free property underlying the network. The reader can refer to Liu et al. (2008) for the detailed analysis of the network properties. In the present work, we utilise the network features to develop a classification scheme to cluster the pockets into small groups to trace the relationships between protein structure and function.

Figure 1 The constructed pocket similarity network: (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 (see online version for colours)



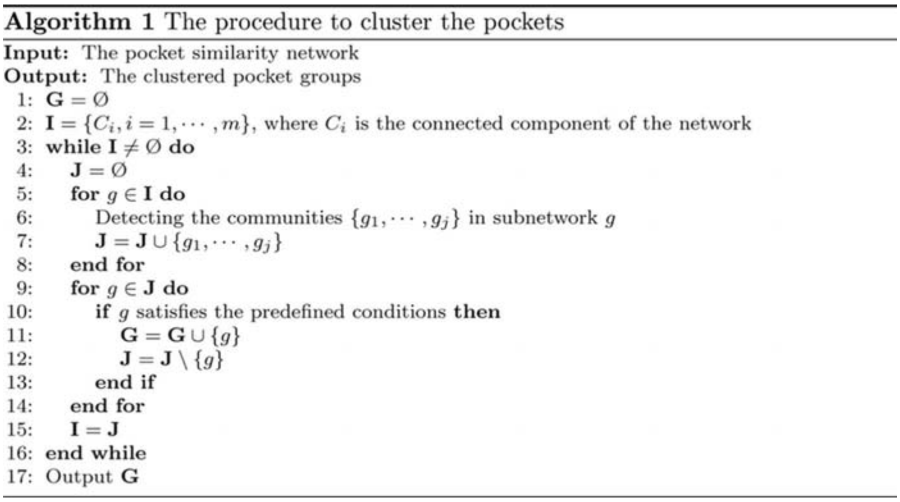
2.2 *Clustering the pockets into small groups*

The community structure feature identified in the pocket similarity network provides us a simple way to cluster the pockets into small groups. The proposed clustering method is based on a well-known concept of modularity Q (Newman, 2004), which is a quality function to measure whether a particular partition of network 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 nodes 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. Generally, Q values for networks typically fall in the range from about 0.3 to 0.7 (Newman, 2004), while a value near 1 indicates strong community structure. Detecting the partition of groups that maximises Q is believed to be a *NP-hard* problem, which makes a brute force exploration impossible for large scale networks with hundreds or thousands of nodes. A fast algorithm to find the approximate optimal partition was proposed in Newman (2004). At the beginning, the algorithm regards every single node as a cluster, then a pair of clusters are merged into one to ensure that their union will produce the biggest increment of modularity Q. Since every step joins one pair of linked clusters to increase Q, there are at most m steps, i.e. the number of edges of the network. The change of Q in each step can be computed in constant time. The process is repeated until only one cluster remains. Obviously, for a network of n nodes, there would be $n-1$ steps for such joining. By following the merging operations, the hierarchy that reveals the community structure can be built. The algorithm is very efficient and widely used in the study of networks (Clauset et al., 2004).

Algorithm 1 shows the procedure to partition the pocket similarity network. The procedure uses the fast algorithm in Newman (2004) to repeatedly split the subnetworks into smaller subnetworks until the predefined criteria are satisfied. In the experiments, two stop criteria are used. The first is that the size of subnetwork is smaller than a given threshold. The second is that the further partition will produce a Q value smaller than a given threshold.



Algorithm 1 The procedure to cluster the pockets

```
Input: The pocket similarity network
Output: The clustered pocket groups
G = ∅
I = {C_i, i = 1, ··· , m}, where C_i is the connected component of the network
while I ≠ ∅ do
    J = ∅
    for g ∈ I do
        Detecting the communities {g_1, ··· , g_j} in subnetwork g
        J = J ∪ {g_1, ··· , g_j}
    end for
    for g ∈ J do
        if g satisfies the predefined conditions then
            G = G ∪ {g}
            J = J \ {g}
        end if
    end for
    I = J
end while
Output G
```

The clustered pocket groups are evaluated both in topology and biology. On the one hand we check the value of the modularity Q which is used as the measure of the divisions. During the process of dividing the network, we record the changing of Q. The bigger modularity Q is, the more obviously we can partition the network into smaller communities. This is considered in the proposed algorithm. When the calculated Q value of potential partition in a particular subnetwork is not significant, the partition is not acceptable and the algorithm stops further dividing the subnetwork. On the other hand, we also analyse the functional consistence and calculate the significant functions in these small groups. The high modularity Q of the clusters and the functional features underlying these groups show that the divided groups are topologically and biologically meaningful.

3 Results

3.1 The clusters of pocket group

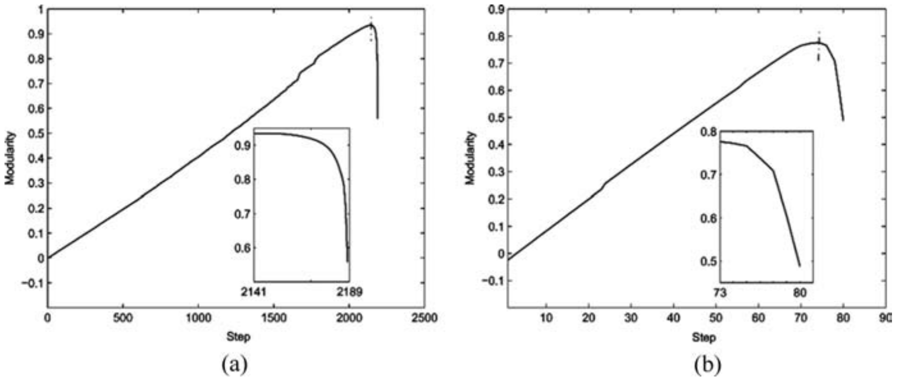
There are 5387 nodes and 4943 edges in the constructed pocket similarity network. From the statistics of the similarity network shown in Figure 1, the network contains 880 connected components. Most of the components contain a few nodes. The maximum connected component of the subnetwork contains 2190 nodes with 2548 edges. The second largest connected component contains 81 nodes with 83 edges. Figure 1 also shows that the similar pockets are naturally clustered together. Since small connected components may contain less information, we take 81 as the threshold of pocket group size, i.e., the dividing process will stop when the size of pocket group is smaller than 81. Therefore in this example we only need to partition the largest two connected components.

Table 1 records the indices 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 2 records the change of Q. The cut-off point of the joining steps with the maximum modularity is also shown. Parts (a) and (b) correspond to the largest and the second largest connected components respectively.

Table 1 The number of the clusters of pockets after first level partition

Connected component	*Node*	*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	–

Figure 2 The changing modularity Q with joining of nodes to clusters: (a) the largest connected component and (b) the second largest connected component



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 partition 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	*Node*	*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 partition of 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 communities. Figure 3(a) records the change of Q and the cut-off point (Step 106) of the hierarchy tree. Figure 3(b) shows the derived 12 smaller communities and the structural feature in one of the communities, which is approximately the common structure among the 9 similar pockets in the same group. In the similar way, the second largest cluster is divided into 10 smaller communities (the cut-off point is Step 86). Figure 3(c) and (d) show the results.

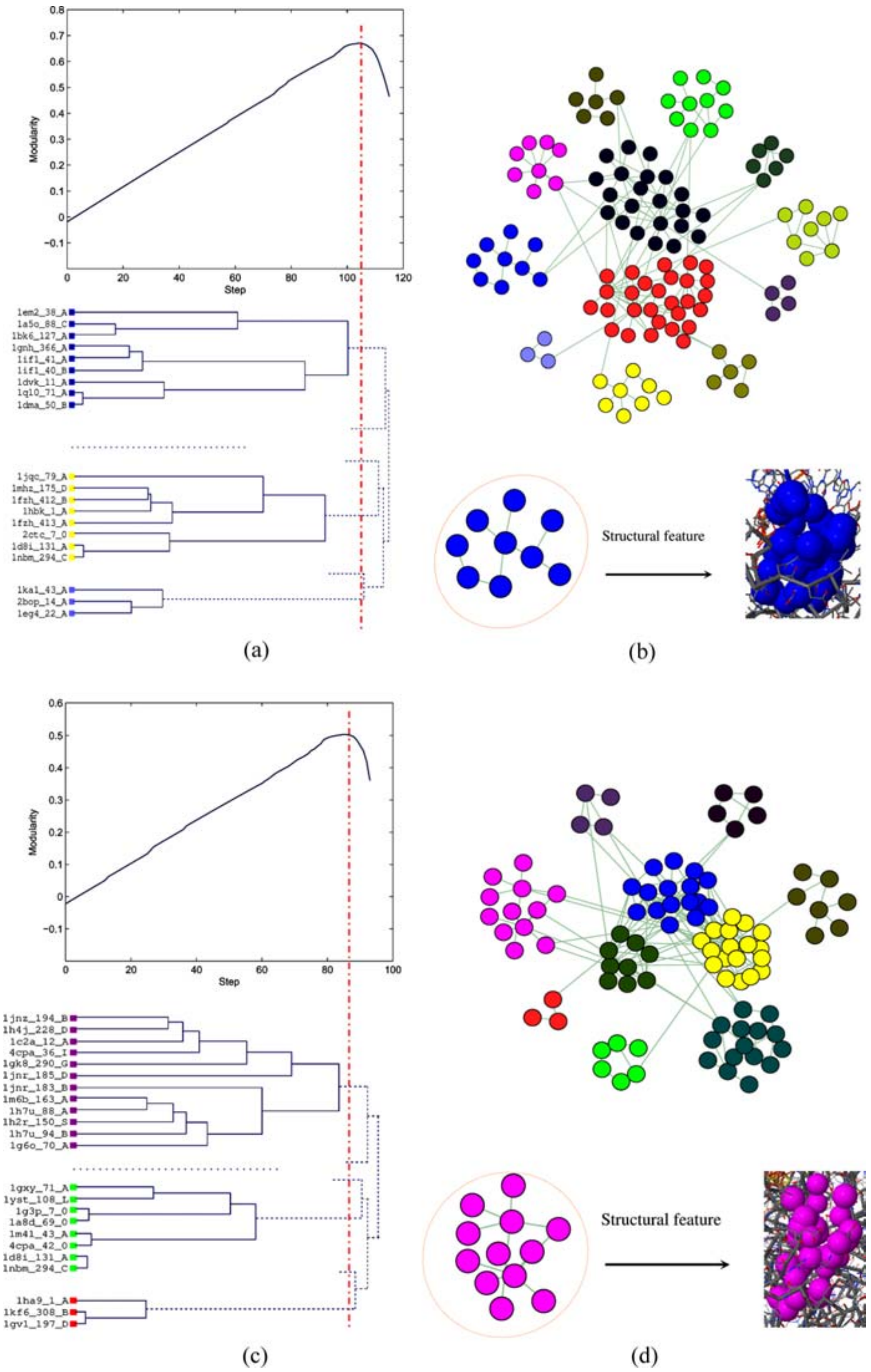
To describe the partition in the Figure 3 more clearly, we plot the clusters in the similarity matrix. The results are shown in Figure 4. The rectangular modules with green color in the figure are correspond to pocket groups respectively. The red points mean that two corresponding pockets are structurally similar, while the blue points indicate that there are no similarity. As shown in Figure 4, most of the red points are grouped together and located in rectangular modules, i.e., the pocket groups, which demonstrates the high clustering quality of the pocket groups.

Totally, the pocket similarity network is partitioned into $(878 + 47 + 8 + 12 + 10) = 955$ clusters, which are regarded as pocket groups. The maximum connected component is divided to 69 clusters. Of course, we can change the threshold of the maximal size of the clusters. The number of pocket groups will increase if a smaller value is chosen. For instance, if we choose 37 as the threshold, the similarity network would be partitioned into 1137 small groups.

3.2 Functional features lie 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 proteins in which the pockets are located. In the 955 pocket groups, there are 816 (85.45%) groups in which at least two pockets have GO terms, i.e., the proteins containing the pockets have GO annotations in GOA database (The Gene Ontology Consortium, 2000). 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 significantly common (2/3) GO terms in their GOA part, i.e., 2/3 of the annotated pockets of the GOA part have common GO terms. And there are 578 (70.83%) pocket groups with significantly common (1/2)

Figure 3 The changing modularity Q with the step of joining nodes to groups and the derived groups in the largest connected component: (a) changing modularity Q in the largest cluster; (b) the divided groups of the largest cluster with the structural feature in one of the groups; (c) and (d) are the results of the second largest cluster (see online version for colours)



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 group provides more evidences: the similar pockets have similar functions, and the pocket groups are functionally important structural motifs for proteins and potentially are the bridges of protein structure and protein function.

Figure 4 The pseudo-colour matrices of the clustering on the top two largest subnetworks after the first level partition in the maximum connected component. Most pairs of similar pockets group in the rectangular modules (see online version for colours)

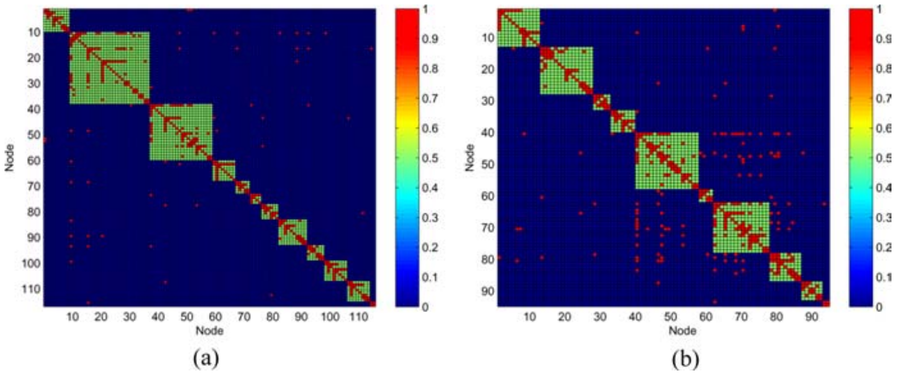


Table 3 The functional similarity among the pocket groups. We annotate the pockets by the GO terms of the proteins containing them. '–' indicates the value that we need not calculate

Group size	*Number*	*GOA status (size of GOA part : number of groups)*	*Common GO terms*					*Percentage*
			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	–	–	–	–	–	

To assess the functional significance in the obtained pocket groups, we measure the enrichment of similarity by an accumulative hypergeometric test (Hwang et al., 2006). The p-value is a probability that a cluster is enriched with a particular function by chance alone. It is defined as:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i}\binom{n-f}{m-i}}{\binom{n}{m}},$$

where n is the number of nodes in the network, f is the number of pockets in the network annotated with a particular GO function, m is the group size and k is the frequency of the GO term in one group. We use a recommended threshold of 0.05 for all validations (Hwang et al., 2006). Smaller p-value indicates that pocket groups are significantly enriched for the specific GO function and can be considered to be functional modules with high probability. We also calculate the density of every community measured by $D_s = 2e/m(m+1)$, where m is the number of the pockets in the same group, i.e., the size of the group, and e is the number of edges. In the 955 clusters, 942 clusters have at least one GO annotation up to now. 341 (36.20%) groups have the significant functions. When we omit the 455 groups containing only 2 nodes, there are 500 groups left, in which 286 (67.14%) groups have the significant GO functions in the 426 groups with at least one GO annotations. Table 4 shows the results of 5 clusters randomly chosen with different sizes (the full list is available upon request). The first column is the group identifier. 'Size' is the number of pockets in each group. 'Density' indicates the densities of the groups. 'H' is the percentage of the pockets consistent with the major functions of the 'GO term' column with the highest statistical significance in the 'Minimum p-value' column in the group. 'D' is the percentage of pockets discordant with the major functions and 'U' is the percentage of pockets not annotated any GO functions. Moreover, we also analyse other GO functions with statistical significance in every pocket group (results are not listed here). In this way, we can get the potentially functional annotations to each pocket group.

Table 4 The significant GO functions in some groups

Group			*Distribution*			*Minimum*	*GO term*
ID	*Size*	*Density*	*H*	*D*	*U*	*p-value*	*and its description*
1	43	0.049	0.233	0.628	0.140	1.53×10^{-13}	GO:0008168: methyltransferase activity
2	23	0.115	0.435	0.478	0.087	1.11×10^{-16}	GO:0008289: lipid binding
3	20	0.100	0.200	0.600	0.200	3.14×10^{-4}	GO:0004601: peroxidase activity
4	9	0.333	0.556	0.333	0.111	7.23×10^{-8}	GO:0008863: formate dehydrogenase activity
5	4	0.500	1.000	0.000	0.000	2.09×10^{-3}	GO:0003824: catalytic activity

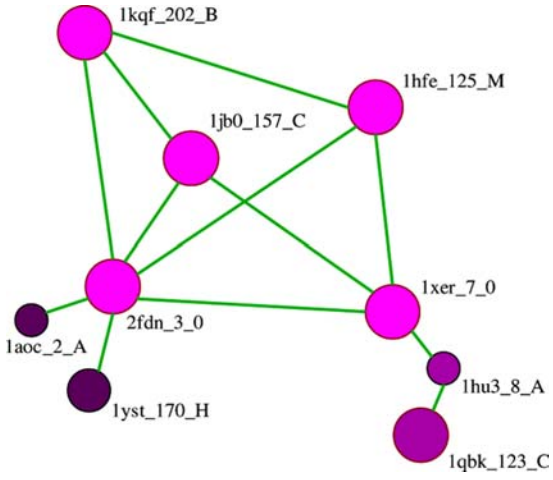
Table 5 An example of the pocket group. The pockets are demonstrated with their sequences and structure features respectively

Pocket	*Len*	*Vol*	*Protein*	*Deg*	*Sequence*
1kqf_202_B	14	184.96	Oxidoreductase, formate dehydrogenase n from E. coli	3	HIEGGLAAEWRAKT
1jb0_157_C	14	177.16	Photosynthesis, crystal structure of photosystem i: a photosynthetic reaction center and core antenna system from cyanobacteria	3	VCPTVLCVGCKRCV
2fdn_3_0	14	173.50	Electron transport, ferredoxin from clostridium acidi-urici	6	Y*C*PVAII*C*ID*C*GA*C*
1yst_170_H	9	83.90	Photosynthetic reaction center	1	FTRASDCGA
1aoc_2_A	4	16.21	Coagulation factor, Japanese horseshoe crab coagulogen	1	CVDC
1hfe_125_M	14	191.58	Hydrogenase, a resolution structure of the Fe- only hydrogenase from desulfovibrio desulfuricans	3	V*C*PTAII*C*IN*C*GQ
1xer_7_0	14	197.04	Electron transport, structure of ferredoxin	3	V*C*PVVF*C*IF*C*MA*C*V
1hu3_8_A	5	9.50	Translation, middle domain of human eif4gii	2	QFLAN
1qbk_123_C	13	192.96	Nuclear transport protein complex structure of the karyopherin beta2-ran gppnhp nuclear transport complex	1	LV*GTGK*FIIFC*N*I

3.3 One pocket group: a case study

To identify the clusters with both structural similarity and functional similarity among the pockets, we study a specific pocket group. Table 5 lists the 9 pockets in the 4th group of Table 4, which locate on proteins of different families. The linkages between pockets in the group are shown in Figure 5 to represent the structural similarities between pockets. In Table 5, 'Vol' column is the volume of the pocket, 'Len' column is the number of amino acids involving in the pocket, and 'Deg' is the number of edges linked to the pocket. As shown in the figure and table, the importance of each pocket to the group's comprehensive shape can be reflected from the degree of the pocket. The pocket with highest degree is 2fdn_3_0 (PDB ID, Pocket ID and Chain ID). We also concatenate the amino acid residues on the primary sequence to constitute the pocket sequence. There are some annotated functional sites (the italic characters in the sequences in Table 5) in Uniprot database. We can find the similarity among the sequences of the pockets, especially among those with higher degree. Table 5 shows the implications that the structure features and sequence features of the particular group are determined by the pockets with high links, and the sequence and structure consistency among the pockets in the same group would determine similar functions. The top three most significant GO functions of the group are GO:0008863 (7.23×10^{-8}, formate dehydrogenase activity, F), GO:0005737 (2.13×10^{-4}, cytoplasm, C) and GO:0006118 (4.67×10^{-3}, electron transport, P). The three annotations in the bracket mean the p-value, the description and the ontology of the GO term respectively. The functional significance in the group provides evidences that the shape features of the pocket groups would imply the similar functions. From the pocket sequences, which are the amino acid residues concatenated from the primary sequence, we can find that the shape of pockets constitutes a spatial profile to perform certain functions and then the common structure features in the group are the functional motifs. Moreover, the pockets are located on diverse surfaces of proteins which come from different families. From the similarity of the structural and functional features of the pockets in the same group, we can get more information about the evolution among the proteins. This pocket group and the above analysis show that the similarity among regionally spatial structures would determine the similar function of proteins, and the surface motifs are crucial to protein function.

Figure 5 The topology of linkage among the pockets in one clustered group (see online version for colours)



4 Discussion and conclusion

Biological functions of a protein are carried out through interacting and binding other molecules on the protein's surface regions. The surface always contains many pockets which have shown high relevance to active sites. The classification of these patterns provide valuable insights into the relationship between protein surfaces and functions.

In this paper, we proposed a novel method to cluster the pockets to small groups based on the feature of community structure of the similarity network. We also provided measurements to the clustering scheme in terms of the modularity Q and revealed the implications of functional similarity and significant functions among the pockets in every group, which provide evidences that these pocket groups are the clusters with both structural and functional similarity.

Our clustering method is based on the attribute of the similarity among the pockets. The structural similarity features among the pockets in a database level have been explored by topological properties of the pocket similarity network (Liu et al., 2008). The community structure underlying the similarity network provides implications that these pockets can be clustered in a hierarchical manner. This is an entirely new clustering scheme which stresses importance on the structural similarity among the pockets and it can be categorised as a hierarchical clustering, although there are some other clustering methods, such as K-means (Jain et al., 1999). Directly using the traditional clustering methods may have risk in the structural data of the pockets on protein surface. 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. Moreover, we cut the hierarchy tree at the step that maximises the modularity value Q during the process. The communities of the similarity network can be extracted efficiently at the position and then the number of the clusters are determined naturally. A direct comparison and measurement among these clustering methods is a challenging task for us in the future. Obviously, the method presented in this paper can be easily extended to other objects in protein's universe, which would provide new valuable insight and more contributions to both the structural genomics and systems biology (Chen et al., 2006a, 2006b; Wang et al., 2006).

The structural similarity among these surface patterns provides valuable information to detect the conserved spatial features and functions from the structural perspective. These pocket groups in a database level have important applications in functional genomics. One direction is to develop a library of structural motifs using these pocket groups. The pockets in the same group generally correspond to similar functions. Although it is difficult to determine the particular functions of a pocket, we identified the functional similarity among the proteins containing the pockets in the same groups which provide strong implications of the common functions underlying in every group. The statistical significance of the GO functions can be used as the potentially functional annotations of the groups. When the concrete functions of a group are identified by more advanced techniques, the group is a functional template and might be used for predicting function of proteins whose function cannot be inferred by the classical sequence and/or global structure comparison methods. These functionally important pocket groups also provide structural information to the binding shape on protein surface, which can be used in drug design or other bioengineering. The target sites indicate potential specificity of the binding ligands. 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 works. We divided the network into clusters without overlaps but in fact some pockets can be both merged into more than one group. The overlap pocket would have important implications in functional diversity of pockets. Based on the network model, more advanced clustering method can be developed to implement this task. This is one of our future directions.

From the example of the divided pocket groups, we can find that most of the pockets in the same groups have the similar structural features, even the sequence features. This provides evidences that the clustered pocket groups not only are the communities in the similarity network, but also have biological meanings. If we want to detect the local structure feature of some functional motifs more concisely, few pockets in every group can be filtered by more standards which have strong relationship with function. The degree of the pocket in every clustered pocket can be used as a direct measure. These pocket groups can be polished carefully in this direction as functional motifs. This can be extended to identify local communities directly in the similarity network, which is an alternative way to detect functional motifs corresponding to dense subgraphs, such as the cliques. Analysing the physicochemical features of the pockets and the multiple sequence alignment of these functional motifs can provide valuable information about the different proteins from different families and species (Ma et al., 2003; Tseng and Liang, 2006). Moreover, the pockets are important surface features, and then the functional surface motifs will have important applications in functional genomics (Zhang and Kim, 2003; Orengo et al., 1999; Nayal and Honig, 2006). We would present another paper about these topics.

In conclusion, we developed a novel network-based classification scheme to cluster the pockets into similar groups at a database level. The method is based on the unique features of the similarity network which maps the structural relationship in a systematic way. We modified the community structure detecting algorithm to partition the network into small clusters. Our method belongs to a hierarchical clustering. 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 and the statistical significant enrichment of GO functions show 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

The authors wish to thank Dr. Jie Liang for providing us the data of CASTp and pvSOAR databases. Thanks also due to Dr. Mark Newman for providing us the software of detecting the community. 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. 2006CB503900) from the Ministry of Science and Technology of China. The second author is also supported by the Knowledge Innovation Program of the Chinese Academy of Sciences.

References

Barabasi, A.L. and Albert, R. (1999) 'Emergence of scaling in random networks', *Science*, Vol. 286, pp.509–512.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) 'The protein data bank', *Nucleic Acids Res.*, Vol. 28, pp.235–242.

Binkowski, T.A., Adamian, L. and Liang, J. (2003) 'Inferring functional relationships of proteins from local sequence and spatial surface patterns', *J. Mol. Biol.*, Vol. 332, pp.505–526.

Binkowski, T.A., Naghibzadeh, S. and Liang, J. (2003) 'CASTp: Computed Atlas of Surface Topography of proteins', *Nucleic Acids. Res.*, Vol. 31, pp.3352–3355.

Binkowski, T.A., Freeman, P. and Liang, J. (2004) 'pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins', *Nucleic Acids. Res.*, Vol. 32, pp.W555–W558.

Binkowski, T.A., Joachimiak, A. and Liang, J. (2005) 'Protein surface analysis for function annotation in high-throughput structural genomics pipeline', *Protein Sci.*, Vol. 14, pp.2972–2981.

Chen, L., Wu, L.Y., Wang, Y. and Zhang, X.S. (2006a) 'Inferring protein interactions from experimental data by association probabilistic method', *Proteins*, Vol. 62, pp.833–837.

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

Clauset, A., Newman, M.E. and Moore, C. (2004) 'Finding community structure in very large networks', *Phys. Rev. E.*, Vol. 70, 066111.

Ferre, F., Ausiello, G., Zanzoni, A. and Helmer-Citterich, M. (2004) 'SURFACE: a database of protein surface regions for functional annotation', *Nucleic Acids Res.*, Vol. 32(Database), pp.D240–244.

Girvan, M. and Newman, M.E. (2002) 'Community structure in social and biological networks', *Proc. Natl. Acad. Sci. USA*, Vol. 99, pp.7821–7826.

Greene, L.H. and Higman, V.A. (2003) 'Uncovering network systems within protein structures', *J. Mol. Biol.*, Vol. 334, pp.781–791.

Hobohm, U. and Sander, C. (1992) 'Selection of a representative set of structures from the Brookhaven protein data bank', *Protein Sci.*, Vol. 1, pp.409–417.

Hwang, W., Cho, Y.R., Zhang, A. and Ramanathan, M. (2006) 'A novel functional module detection algorithm for protein-protein interaction networks', *Algorithm for Molecular Biology*, Vol. 1, 24.

Jain, A.K., Murty, M.N. and Flynn, P.J. (1999) 'Data clustering: a review', *ACM Computing Surveys*, Vol. 31, pp.264–323.

Jones, S. and Thornton, J.M. (1997) 'Prediction of protein-protein interaction sites using patch analysis', *J. Mol. Biol.*, Vol. 272, pp.133–143.

Kinoshita, K. and Nakamura, H. (2005) 'Proteins identification of the ligand binding sites on the molecular surface of proteins', *Protein Sci.*, Vol. 14, pp.711–718.

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

Laurie, A.T. and Jackson, R.M. (2005) 'Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites', *Bioinformatics*, Vol. 21, pp.1908–1916.

Liang, J., Edelsbrunner, H. and Woodward, C. (1998) 'Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for Ligand design', *Protein Sci.*, Vol. 7, pp.1884–1897.

Liu, Z-P., Wu, L-Y., Wang, Y., Zhang, X-S. and Chen, L. (2008) 'Analysis of protein surface patterns by pocket similarity network', *Protein and Peptide Letters*, Vol. 15, pp.448–455.

Ma, B., Elkayam, T., Haim, W. and Nussinov, R. (2003) 'Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces', *Proc. Natl. Acad. Sci. USA*, Vol. 100, pp.5772–5777.

Nayal, M. and Honig, B. (2006) 'On the nature of cavities on protein surfaces: application to the identification of drug-binding sites', *Proteins*, Vol. 63, pp.892–906.

Newman, M.E. (2004) 'Fast algorithm for detecting community structure in networks', *Phys. Rev. E.*, Vol. 69, p.066133.

Orengo, C.A., Todd, A.E. and Thornton, J.M. (1999) 'From protein structure to function', *Curr. Opin. Struct. Biol.*, Vol. 9, pp.374–382.

Rao, F. and Caflisch, A. (2004) 'The protein folding network', *J. Mol. Biol.*, Vol. 342, pp.299–306.

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

Stark, A., Sunyaev, S. and Russell, R. (2003) 'A model for statistical significance of local similarities in structure', *J. Mol. Biol.*, Vol. 326, pp.1307–1316.

Strogatz, S.H. (2001) 'Exploring complex networks', *Nature*, Vol. 410, pp.268–276.

The Gene Ontology Consortium (2000) 'Gene ontology: tool for the unification of biology', *Nature Genet.*, Vol. 25, pp.25–29.

Tseng, Y.Y. and Liang, J. (2006) 'Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: a Bayesian Monte Carlo approach', *Mol. Biol. Evol.*, Vol. 23, pp.421–436.

Wang, Y., Joshi, T., Zhang, X.S., Xu, D. and Chen, L. (2006) 'Inferring gene regulatory networks from multiple microarray datasets', *Bioinformatics*, Vol. 22, pp.2413–2420.

Watts, D.J. and Strogatz, S.H. (1998) 'Collective dynamics of small-world networks', *Nature*, Vol. 393, pp.440–442.

Wuchty, S. (2001) 'Scale-free behaviour in protein domain networks', *Mol. Biol. Evol.*, Vol. 18, pp.1694–1702.

Zhang, C. and Kim, S.H. (2003) 'Overview of structural genomics: from structure to function', *Curr. Opin. Chem. Biol.*, Vol. 7, pp.28–32.

Zhang, S., Jin, G., Zhang, X.S. and Chen, L. (2007) 'Discovering functions and revealing mechanisms at molecular level from biological networks', *Proteomics*, Vol. 7, pp.2856–2869.