PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Clustering and visualizing similarity networks of membrane proteins

Geng-Ming Hu, Te-Lun Mai, and Chi-Ming Chen*

Department of Physics, National Taiwan Normal University, Taipei, Taiwan

ABSTRACT

We proposed a fast and unsupervised clustering method, minimum span clustering (MSC), for analyzing the sequence–structure–function relationship of biological networks, and demonstrated its validity in clustering the sequence/structure similarity networks (SSN) of 682 membrane protein (MP) chains. The MSC clustering of MPs based on their sequence information was found to be consistent with their tertiary structures and functions. For the largest seven clusters predicted by MSC, the consistency in chain function within the same cluster is found to be 100%. From analyzing the edge distribution of SSN for MPs, we found a characteristic threshold distance for the boundary between clusters, over which SSN of MPs could be properly clustered by an unsupervised sparsification of the network distance matrix. The clustering results of MPs from both MSC and the unsupervised sparsification methods are consistent with each other, and have high intracluster similarity and low intercluster similarity in sequence, structure, and function. Our study showed a strong sequence–structure–function relationship of MPs. We discussed evidence of convergent evolution of MPs and suggested applications in finding structural similarities and predicting biological functions of MP chains based on their sequence information.

Proteins 2015; 83:1450–1461.

Key words: network clustering; protein similarity networks; sequence homology; protein structure; protein function; membrane proteins.

INTRODUCTION

Over the past two decades, there has been a rapid growth in volume and diversity of biological data, such as new sequence, structural, and functional data added to various proteomic and genomic repositories. To organize these vase amounts of diverse data, many bioinformatics databases have been established and made available to the public. Various statistical methods have been developed for extracting useful knowledge from volumes of data.[1–9] Recently, studies of biological networks have received considerable attentions. For example, they are used to analyze protein interactions, metabolic pathways, and gene regulatory mechanisms.[10–12]

Network clustering is one of the fundamental problems in recognizing patterns of complex networks. Traditionally there are two commonly used approaches for network clustering, including the partitional and the hierarchical algorithms.[13–15] Partitional clustering algorithms, such as *K*-means algorithm,[16] group complex networks into several clusters so that nodes within each cluster are more closely related to each other than nodes in different clusters. However, this method is computationally difficult, requires a priori knowledge about the number of clusters, and is sensitive to the choice of initial cluster centers. On the other hand, hierarchical algorithms, such as the hierarchical clustering (HC) method,[15] group a network with a sequence of nested partitions, either from singleton clusters to a cluster including all individuals or vice versa. HC suffers from the same problems of a predetermined number of clusters and nonunique solutions. Another popular clustering method, graph-based sparsification clustering, starts with a proximity matrix, eliminates long distance edges with a threshold, and preferentially retains the edges that are likely to be part of the same cluster.[17–19] Although it is simple, its drawback is the arbitrary threshold which varies with the system under investigation, as revealed by a recent attempt in studying how dynamic graph thresholding relates to various network clustering approaches,[5]

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: National Science Council of Taiwan;; Grant number: MOST 102-2112-M-003-010-MY3.
*Correspondence to: C.-M. Chen; Department of Physics, National Taiwan Normal University, Taipei, Taiwan. E-mail: cchen@phy.ntnu.edu.tw
Received 17 February 2015; Revised 23 April 2015; Accepted 17 May 2015
Published online 25 May 2015 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24832

including a popular flow-based approach, Markov clustering algorithm (MCL).[20] However, there exists a large ambiguity in this study to the predicted value of threshold, and the clustering results shown in its supplementary document are hardly satisfactory. Therefore, it is desirable to have an unsupervised clustering algorithm that is computationally efficient and gives reliable results.

For the study of protein networks, the small-world network (SWN) approach has been demonstrated to be useful in analyzing their structures and functions.[17,21,22] Furthermore, protein similarity networks (PSNs) have been used to visualize functional trends across protein superfamilies from the context of sequence similarity.[17] Subtle sequence similarities frequently indicate structural, functional, and evolutionary relationships among protein sequences. The use of PSNs to visualize trends of sequence and structure has been made to make functional inferences for the cupin superfamily.[18] A review on proteins as networks can be found in an article by Krishnan *et al.*[23]

Recently, there is an increasing interest in the network analysis of membrane proteins (MPs).[24] MPs play a key role in a wide variety of biological processes; their functions include cell–cell contact, surface recognition, cytoskeleton contact, signaling, enzymatic activity, or transporting substances across the membrane.[25] The biological functions of MPs are strongly related to their three dimensional structures. MPs in the same superfamily usually have similar three-dimensional structures but diverse functions stemming from multiple structural distortions. Many known diseases result from the defects of MPs. Their clinical importance is demonstrated by the fact that [mt]50% of known drugs in use today target MPs, which are also responsible for the uptake, metabolism, and clearance of these pharmacologically active substances. Despite their biological and pharmaceutical importance, due to difficulties in crystallizing MPs, only about 500 unique structures have been derived so far.[26,27] As the attempts of using experimental methods to study MPs have encountered difficulties, there exist great incentives for computational and statistical studies of MPs.

To investigate the sequence–structure relationship of MPs, we have previously shown that the structure of several MPs, including bacteriorhodopsin, halorhodopsin, and sensory rhodopsin, can be predicted from their sequence using thermodynamic principles.[26,28,29] Furthermore, we have combined molecular dynamics simulations and fold identification procedures to predict the structure of 696 kinked and 120 unkinked transmembrane (TM) helices from their sequence in the protein data bank of transmembrane proteins (PDBTM).[30,31] Although these results delineate a tight correlation between sequence and structure of MPs, a panoramic view of their sequence–structure–function relationship is desirable. The aim of the present work is to develop an integrated approach for clustering and visualizing PSNs for MPs and to investigate their sequence–structure–function relationship. Here we propose the use of minimum span clustering (MSC) algorithm,[4] which is efficient (the computation time is linear with system size) and does not require predetermined inputs on the number or size of clusters. Moreover, MSC enables the users to view a complex system at various characteristic resolutions. Our earlier application of MSC has provided an excellent clustering of the social science network consisting of 1575 SSCI (Social Sciences Citation Index) journals at four characteristic resolutions.[4]

In the present study of PSNs, the general structure of the sequence similarity network (SSN) of 682 MP chains was investigated by both MSC and the minimum spanning tree (MST) algorithms.[7,32] Our methods provide an efficient and convenient approach to observe the sequence–structure–function relationship among large sets of MPs, to identify a general sequence pattern for a group of structure/function related proteins, and to investigate the similarities/differences between neighboring protein groups in the network. A characteristic threshold distance for the cluster boundary of SSN for MPs was identified from analyzing the statistical edge distributions in the SSN, which was used to view the network connectivity of MPs in an unsupervised sparsification clustering. These methods of network clustering are described in "Methods", followed by results and discussion in "Results and Discussion". In "Conclusion", we conclude a strong sequence–structure–function relationship for MPs, and the feasibility of predicting structure and function of MPs based on their sequence information.

METHODS

Dataset construction

The sequence and structure dataset of 682 MP chains used in this study was downloaded from PDBTM (http://pdbtm.enzim.hu/). Sets of protein sequences from PDBTM were sifted with a protein sequence culling server, PISCES,[33] by criteria of sequence identity (pairwise sequence identity <95%) and structural quality (X-ray crystal resolution better than 4 Å and the traditional crystallographic R-factor better than 0.36). For protein pairs with sequence identity [mt]95%, only the highest resolution polypeptide chain was considered. Sequences consisting of noncanonical amino acids were excluded in this study. The determination of transmembrane region of a chain was made according to the TMDET and OPM databases. In our dataset, 176 chains are peripheral and 506 chains have transmembrane units. The functions of MP chains were assigned and classified according to the Pfam descriptions and molecular functions in gene ontology (GO). Overall, 627 chains are classified into 42 functional groups, 15 chains have unique function, and 40 chains have no characterized functions. To visualize these functional groups, chains in the first 24 groups were colored by their function (similar colors for similar functions). Chains have no characterized functions were represented by open circles. All other chains were colored in black. For simplicity in the color representation of protein groups, the description of molecular functions is more specific for large groups, and is more general for small groups. A detailed list of our dataset is available in Table SI of the supporting information.

Distance/similarity matrix calculation

To calculate similarity between sequences in our dataset, we considered the E value calculated from BLAST (Basic Local Alignment Search Tool), a parameter that describes the number of hits one expected to see just by chance when searching the database of a particular size.[5,17,34] This E value was calculated using the general scoring matrix BLOSUM62 with default parameters, and its cutoff was chosen to be 10^8 for a comprehensive coverage. Such an E value (or $-\log_{10} E$) between sequences has been widely used as a distance (or similarity) measure to visualize the clustering of protein functions in previous studies of protein networks.[5,17,18] Since a lower E value infers a more significant match, here we defined the similarity between sequences i and j as $s_{i,j} = 1/\left(1+E_{i,j}\right)$ with a value between 0 and 1. The similarity pattern of sequence i in the dataset was denoted as $\vec{s}_i = \left\{s_{i,j}\right\},\ j = 1\ \cdots\ 682$. Furthermore, we defined the overall similarity between sequences i and j as the cosine measure of their similarity patterns in the dataset; that is, $\tilde{s}_{i,j} \equiv \frac{\vec{s}_i \cdot \vec{s}_j}{|\vec{s}_i||\vec{s}_j|} = \sum_k s_{i,k}s_{j,k} / \sqrt{\sum_k {s_{i,k}}^2 \sum_k {s_{j,k}}^2}$, where k spans all dataset of 682 chains. Finally, we defined the distance between sequences i and j in the distance matrix $\{d_{i,j}\}$ of our dataset as

$$d_{i,j} = 1/\tilde{s}_{i,j} - 1. \tag{1}$$

The tertiary structure similarity of MPs was calculated with TM align, a protein structure alignment algorithm based on the TM score.[35] The transformation between similarity (measured by TM score) and distance of tertiary structure was also defined as Eq. (1). TM-align exploits three kinds of quickly identified initial structural alignments, including an alignment of secondary structures between two proteins using dynamic programming (DP), a gapless matching alignment of two structures, and another DP alignment with a mixed score matrix of the above two alignments and a gap-opening penalty. These initial alignments are then submitted to a heuristic iterative algorithm, which has been extensively used in refining NP-hard structure-based alignments. It has been found that the posterior probability from various datasets has a similar rapid phase transition at TM score about 0.5, suggesting that protein pairs with a TM score > 0.5 are mostly in the same fold while those with a TM score < 0.5 are mainly not in the same fold. An all-against-all structure alignment of 10,515 nonredundant protein chains in the PDB has been conducted and about 2000 folds has been obtained after clustering all structures using the threshold of TM score = 0.5.[36]

Network visualization

In general, the comprehensive structure of a biological network is rather complex and it is difficult to recognize its important features from a hairball diagram. To extract the characteristics of the network, in this study we applied two visualization methods, including MST and an unsupervised sparsification clustering. There are a number of algorithms to construct MSTs, and we used the Kruskal algorithm[32] in this study. Decision about whether to connect a pair of sequences of the seeding graph was made using the distance array $d_{i,j}$, which was sorted in the order of increasing distance. The sorted distance array $d_{i,j}$ was scanned from its top and a linkage between two sequences was added to the seeding graph if no loop was present. At the end of this procedure, a complete MST was constructed for the SSN (or for a protein cluster). With the network information provided in the above distance matrix, MST shows a tree diagram of the network with the shortest path connecting all nodes in the network. On the other hand, the unsupervised sparsification clustering method reads the characteristic threshold distance, D_t, and constructs a network graph by Cytoscape,[37] showing all connected edges with a distance shorter than D_t. Since D_t is set to be the characteristic distance of the network separating intracluster edge distribution from intercluster edge distribution from our MSC analysis (as described in the "Results and Discussion" section), each protein cluster of the network and its intracluster connectivity can be clearly visualized in the Cytoscape graph. We note that, for both visualization methods, the relative positions of nodes in the graphs do not preserve their actual distance as calculated in the distance matrix.

Network clustering

Network topology is the arrangement of various elements of a network, and clustering of a network helps one to identify the topological structure of the network. Our proposed clustering method, MSC, attempts to cluster the network such that the span is minimized for both the network and each cluster. The constructed tree diagram of the network by MST was further clustered into a tree diagram of clusters by MSC. In this way, as will be discussed in the "Results and Discussion" section, it is feasible to statistically analyze the inter- and intracluster edge distributions of the network and define a threshold distance D_t for the boundary between clusters. Here we

Table I
Shortest Distance Pairs of Network Nodes Listed in Increasing Order for Demonstration Purposes

Node i	3	4	2	1	8	6	7	10	9	5
Node j	4	3	3	8	1	7	6	3	4	8
d_{ij}	0.08	0.08	0.15	0.49	0.49	0.68	0.68	0.90	0.97	0.98

illustrate the procedure of MSC using a simple ten-component three-step example:

We identify the closest neighbor of each node and record their distances in a list of ascending order, as shown in Table I. This list of shortest distances between node pairs is truncated at the threshold distance D_t. For a network of N nodes, instead of dealing with a distance matrix of N^2 elements, MSC only needs to process at most N distances in this list.

The first cluster is constructed by starting from the node pair with the shortest distance, then including additional pairs from the list in the order of increasing distance. For the added distance, if one of the two nodes is involved in one of the above constructed clusters, the size of this cluster increases but the number of clusters remains the same. If both nodes of the distance are not involved in the above constructed clusters, a new cluster is identified and the number of clusters increases. All clusters of the network are found when all distances in the list are considered. In this example, sequences 3 and 4 form the first cluster. This cluster grows by the inclusion of sequence 2 through its connection with sequence 3. Sequences 1 and 8 then form the second cluster, and sequences 6 and 7 form the third cluster. Sequences 10 and 9 are added to the first cluster respectively through their connection to sequences 3 and 4. Finally, sequence 5 is added to the second cluster through its connection to sequence 8. The identified clusters in the first run are referred as the first level clustering, which has the highest resolution.

Clusters constructed in step 2 are considered as renormalized nodes and the average distance matrix of these clusters is calculated for all inter-cluster node pairs between two clusters. The network consisting of these renormalized components is further clustered by steps 1 and 2, and higher levels of clustering with lower resolutions are constructed.

Similarity measure for sets of clustering results

Consider two sets of clustering results, $\mathbf{A} = \{A_1, A_2, \ldots, A_m\}$ and $\mathbf{B} = \{B_1, B_2, \ldots, B_n\}$, of the same dataset. The similarity matrix of **A** and **B** can be expressed as:

$$\mathbf{S}_{\mathbf{A},\mathbf{B}} = \begin{bmatrix} S_{11} & \cdots & S_{1n} \\ \vdots & \ddots & \vdots \\ S_{m1} & \cdots & S_{mn} \end{bmatrix}, \tag{2}$$

where the matrix element $S_{ij} = p/q$ is the Jaccard's similarity coefficient with p being the size of intersection and q being the size of the union of cluster sets A_i and B_j.[38] In this study, the similarity of clustering results **A** and **B** is defined as $\mathrm{Sim}(\mathbf{A}, \mathbf{B}) = \sum_{i \le m,\ j \le n} S_{ij} / \max(m, n)$, and it has been shown that $0 < \mathrm{Sim}(\mathbf{A}, \mathbf{B}) \le 1$ and $\mathrm{Sim}(\mathbf{A},\mathbf{A}) = 1$.[39]

RESULTS AND DISCUSSION

The E value calculated by the homology searching algorithm BLAST has been widely used in recognizing remote protein homologies for the structural and functional annotation of newly determined proteins, and proven to provide useful clustering results for various protein networks. Typically the threshold E value used in clustering proteins is much <1, but there is no general rule for determining the threshold value. Since large E values have little meaning, the defined sequence distance between proteins in Eq. (1) is mainly contributed from those small E value pairs in the overall sequence similarity pattern, as described in the "Methods". Moreover, many protein pairs have very similar structure and function even though their E value is large. By sifting our data set with criteria E value > 1 (or $s_{i,j} < 0.5$) and $\tilde{s}_{i,j} > 0.85$, 96% of protein pairs have similar tertiary structures (TM score > 0.5). Furthermore, we calculated the distribution of sequence similarity for protein pairs in our dataset using the Pfam classification. For related protein pairs (within the same family and having the same Pfam description), 8% (9%) of pairs have a $s_{i,j}$ value less than 0.2 (0.5), but only 4% (6%) of pairs have a $\tilde{s}_{i,j}$ value less than 0.2 (0.5). Furthermore, 99% of unrelated protein pairs have a $s_{i,j}$ or $\tilde{s}_{i,j}$ value less than 0.2. The above findings suggest that the defined overall similarity, $\tilde{s}_{i,j}$, in this study is a better index for homology modeling of proteins, although our clustering results using $s_{i,j}$ or $\tilde{s}_{i,j}$ are in general consistent with the Pfam classification.

To determine the threshold distance (D_t) for cluster boundary, as shown in Figure 1, we compared the length distribution of edges in the MST diagram and that in the 1st level MSC clusters without thresholding. As described in the "Methods" section, MST generates a tree diagram for the network by minimizing its overall span, while MSC divides the network into clusters by minimizing the overall span of each cluster and the sum of overall span for all clusters. In other words, the MST diagram can be decomposed into MSC clusters by removing all intercluster edges. Thus, the difference between the length

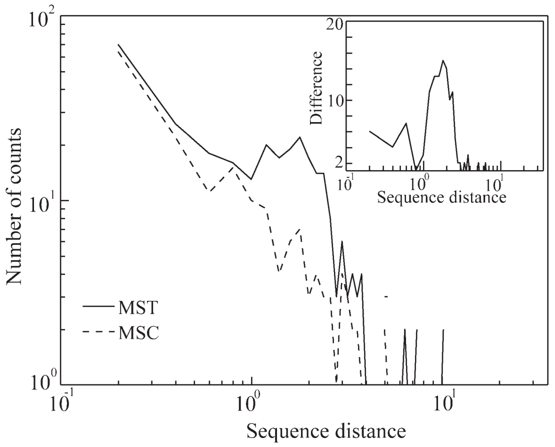

Figure 1

Statistical length distributions of edges predicted by MSC (dashed line) and MST (solid line). As shown in the inset, the difference of these two distributions gives the length distribution of intercluster edges, since the MST edges are composed of intracluster edges (MSC edges) and intercluster edges.

distribution of MST edges and that of MSC edges represents the length distribution of inter-cluster edges, which has a smaller peak at sequence distances less than 1 and a major peak at sequence distances greater than 1, as shown in the inset of Figure 1. The first peak is mostly contributed from clusters with similar sequences, structures, and functions, and these clusters tend to form a group in a lower resolution MSC. On the other hand, the second peak is contributed from clusters with different sequences, structures, or functions. Therefore, this characteristic sequence distance can be set as the threshold distance for the cluster boundary of MPs, and we have $D_t = 1$. This characteristic threshold distance will be used in visualizing the clustering of proteins and the connectivity of clusters in MSC or an unsupervised sparsification clustering. To study the sensitivity of our clustering results on the value of D_t, we calculated the Jaccard's similarity coefficient for clustering results of membrane proteins using different D_t, as shown in Supporting Information Figure S1. The Jaccard's similarity coefficient between the MSC results is greater than 0.9 if the value of D_t is varied between 0.6 and 1.6, and the coefficient between the sparsification clustering results is greater than 0.9 if D_t is varied between 0.7 and 1.3. For other values of D_t, the clustering results would have more significant changes, particularly for using the sparsification clustering method. The value of D_t differs for four different systems that we have studied (MPs, proteases, kinases, and phosphatases), and is different for different levels of MSC clustering. This result is consistent with findings in previous studies.[5]

To cluster MPs, we first applied MSC to decompose their SSN at various resolutions. At the highest resolution (1st level) with $D_t = 1$, the network was decomposed into 156 clusters and 93 isolated nodes. As shown in Figure 2, we plotted a MST diagram for each of these clusters. For the largest seven clusters, we displayed the PDB code of their hubs, including antibody Fab fragments (which were mainly added to stabilize MPs and aid crystallization in experiments), GPCRs, porins, ATP-binding cassette transporters, aquaporins, potassium/sodium channels, and bacterial rhodopsins. Their representative tertiary structures are displayed in Supporting Information Figure S2. Here a node is denoted as a hub if it has more than five connections in a MST diagram. The validity of our MSC clustering is demonstrated by the high average intracluster sequence similarity of 0.89 and the low average inter-cluster sequence similarity of 0.02. In Figure 2, each node is colored according to its function, and the observed color consistency within clusters implies a close sequence-function relationship for MPs. Furthermore, the average intra- and intercluster tertiary structure similarities of our MSC sequence clustering are 0.71 and 0.23, respectively. The percentage of inconsistent links (two proteins that are connected in the sequence network but belong to different functional categories) in Figure 2 is less than 5%, suggesting a strong sequence–function relationship for MPs. As shown in Supporting Information Figure S3, all of these protein pairs have a low TM score about 0.2. In addition, the *E* value is usually much smaller than 1 for consistent links, but is between 0.01 and 1 for most of inconsistent links. The clustering result of proteins can be further improved if their structure information is input as additional information. In Supporting Information Table SII, we listed a detailed data analysis for protein pairs of these inconsistent links. For some inconsistent links, the functions of proteins are related or similar to each other. For example, 3mk7-C is a cytochrome *c* oxidase and 1zrt-D is an ubiquinol–cytochrome *c* reductase. Similarly 3ag3-D is a cytochrome *c* oxidase and 2bs2-C is an oxidoreductase. Moreover, 3vmt-A is a transglycosylase, while 3fwl-A is both a penicillin binding protein and a transglycosylase, suggesting the effect of divergent evolution observed in the MSC result. In our clustering scheme, no protein can be assigned to be in more than one cluster, but in reality proteins can have more than one function. In Table II, we further elaborated the sequence–structure–function relationship for the largest 7 clusters. These results verify a strong sequence–structure relationship for MPs. For comparison, in a random clustering of 682 nodes into 249 clusters, the average intra-/intercluster similarities are 0.03/0.03 in sequence and 0.23/0.23 in tertiary structure. The scatter plot in Supporting Information Figure S4 shows the relationship between the standard deviation of sequence lengths (σ_l) and the average sequence length (l_{av}) for predicted protein clusters. This result indicates that sequence length plays some role in the clustering of proteins, since 85% of clusters have a value of σ_l/l_{av} less

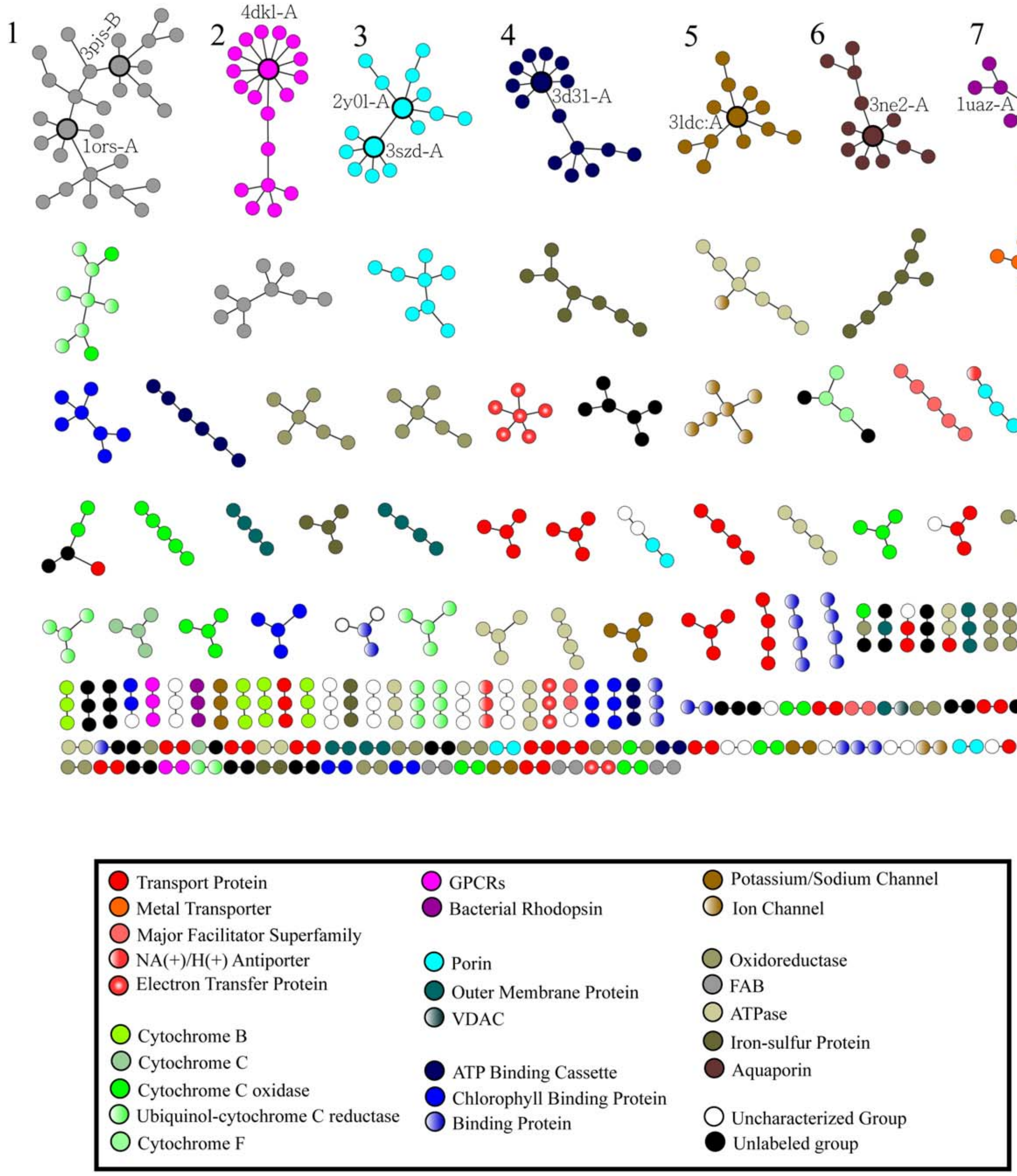

Figure 2

MST diagrams of 156 MSC clusters in the 1st level. These clusters are arranged from left to right in the descending order of cluster size. Here each node is colored according to its function, as described in the legend (only for the largest 25 functional categories to avoid confusion, since many colors look alike). The largest seven clusters are labeled and their hub centers are circled. Isolated nodes are not shown.

than 0.25. The details of our MSC results for MPs can be found in Table SI of the supporting information.

At the second level of MSC with a renormalized threshold $D_t^R = 4$ (as described in step 3 of MSC), the SSN of MPs contains 22 clusters, as demonstrated in the tree diagram of Figure 3. Here solid lines represent intracluster links, springs represent intercluster links, and dashed circles enclose nodes in the same group. The average intracluster similarity is 0.58 in sequence and 0.50 in tertiary structure, while the average intercluster similarity is 0.01 in sequence and 0.23 in tertiary structure. In Figure 3, we observed that clusters in FAB, GPCR, metal transporter, electron transfer protein, and potassium/sodium channel groups tend to form an aggregation of the same function, indicating a close evolutional relationship for chains in the same functional

Table II

Summary of the Sequence, Structure, and Function Properties for the Largest Seven Clusters Predicted by the 1st Level MSC Sequence Clustering of MPs

Cluster	Sequence similarity	Structure (similarity)	Function	Consistency
1	0.97	β sheets (0.70)	Antibody Fab fragments	28/28
2	0.91	7−TM helix bundle (0.63)	GPCRs	19/19
3	0.96	β barrel (0.91)	Porins	15/15
4	0.96	Mix of helices and sheets (0.57)	ATP-binding cassette (ABC) transporter	15/15
5	0.87	α helix bundle (0.50)	Aquaporins	12/12
6	0.96	α helices (0.86)	Potassium/sodium channels	12/12
7	0.94	7−TM helix bundle (0.81)	Bacterial rhodopsins	11/11

group. These MP clusters also have very similar structures, judged by the fact that the average tertiary structure similarity is 0.73 for the FAB group, 0.64 for the GPCR group, and 0.78 for the metal transporter group. Within each group, MPs are phylogenetically related and share conserved sequence patterns. It is interesting to note a possible radiative evolution (instead of bifurcated evolution) of GPCRs from the hub node 4dkl-A, as observed in the sequence similarity diagram of GPCRs in Figure 2. Such a covarion process of GPCRs deserves further investigation using metric multidimensional scaling analysis to explore their sequence space and identify various evolutionary pathways.[40] On the other hand, evidence of convergent evolution is observed for chlorophyll binding proteins (CBPs) and porins. Chains in the CBP group (including MSC clusters 15, 43, 87, 89, 90, 146, 148, and 154) are very distributed (low sequence similarity), and the average tertiary structure similarity for these chains is only 0.22. For simplicity, we focused on the CBPs in clusters (15, 148) and (89, 90). According to the literature, clusters 15 and 148 contain CBPs in plants, while clusters 89 and 90 contain CBPs in photosynthetic bacteria. The average sequence distance is 0.05 between clusters 15 and 148, 0.08 between clusters 89 and 90, and greater than 40 otherwise. Therefore, although the core complexes of photosystems I and II are highly conserved among oxygenic photosynthetic organisms due to their common origin from an endosymbiosis event,[41] CBPs show larger variability in their origin among different groups of organisms, possibly correlating with the

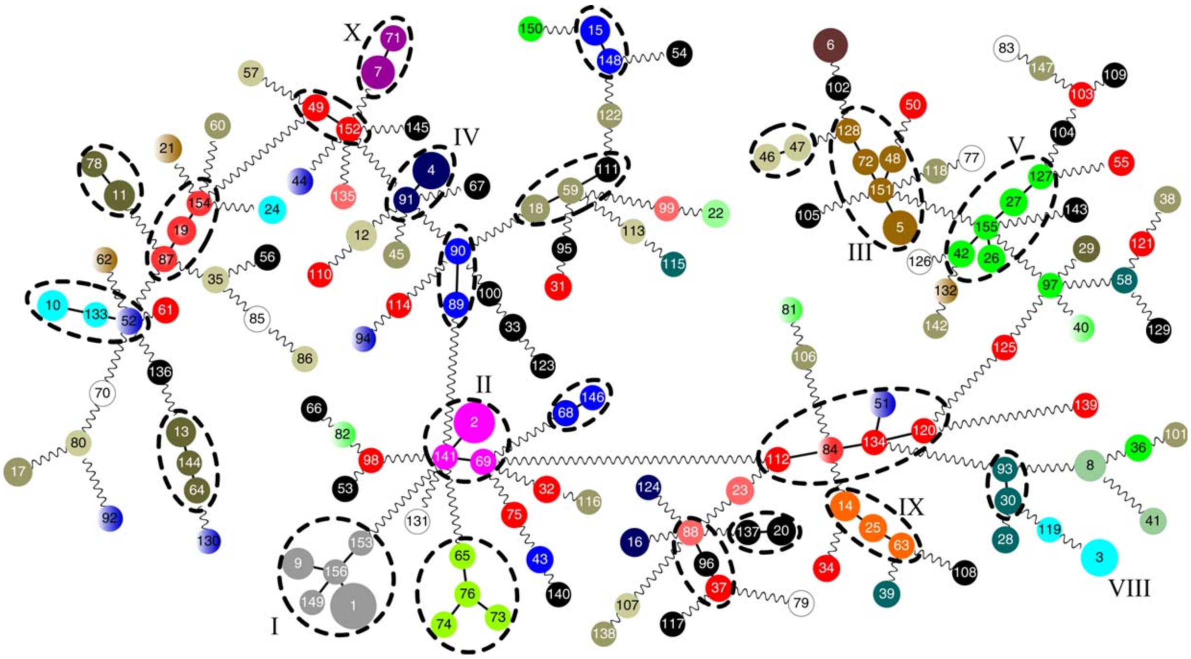

Figure 3

MST diagram of 22 MSC groups in the second level. Intragroup and intergroup links are represented by solid lines and springs, respectively. The size of nodes is proportional to s_c + 25, where s_c is the cluster size. Here each node represents a cluster in the 1st level, and is colored according to the majority function of its components. The color code of nodes is the same as that in Figure 2. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table III

Comparison of Clustering Validity (Intra-/Intercluster Similarity) of the 1st Level MSC, MCL, *K*-means, and HC

	Intra-/intercluster similarity	
Clustering method	Sequence	Tertiary structure
MSC	0.89/0.02	0.71/0.23
MCL	0.60/0.02	0.50/0.23
K-means	0.37/0.03	0.45/0.23
HC	0.89/0.02	0.67/0.23

adaption to their environment. In Supporting Information Table SIII, we randomly selected 77 CBPs of plants, alga, and euglena from the UniProt database and found that their sequences are most similar to CBPs in clusters 15 and 148. Although most of these CBPs are highly homologous, the CBP (UniProt ID: P84988) of Euphrates poplar is distinct from other CBPs, suggesting a different origin of this protein.

In this article, we have mentioned two features of MSC for network clustering, including its computational efficiency and its prediction of the characteristic number of clusters at various resolutions. By running Matlab codes of various clustering methods on an Intel Core i7 desktop computer, we found that MSC is comparable with HC in efficiency, and is much faster than K-means and affinity propagation (AP).[6] For a network of size N = 3,000, the computing time is approximately 2–5 s for MSC and HC, 30 min for K-means, and 8.5 h for AP. For N = 10,000, the computing time is less than 1 min for MSC and HC but longer than 12 h for K-means and AP. To further validate MSC, as shown in Table III, we compared the intra- and intercluster similarity in sequence and in tertiary structure, calculated by various clustering methods (MSC, MCL, *K*-means, and HC). For SSN, the intra-/intercluster similarities are 0.89/0.02 in MSC, 0.60/0.02 in MCL, 0.37/0.03 in *K*-means, 0.89/0.02 in HC, and 0.03/0.03 in a random clustering. For the clustering of tertiary structures, the intra-/intercluster similarities are 0.71/0.23 in MSC, 0.50/0.23 in MCL, 0.45/0.23 in K-means, 0.67/0.23 in HC, and 0.23/0.23 in a random clustering. Here, samples in a random clustering were constructed by randomly assigning 682 nodes into 249 clusters. The intercluster similarity in tertiary structure of all four methods is about the same as that of a random clustering, suggesting that the overall tertiary structures are very different for proteins in different clusters as predicted from the four methods. The largest intracluster similarities for MCL clustering were found with an inflation of 18.7, although the conceivable values of inflation are between 1.1 and 10.0. In general all four methods provide much better clustering results than a random assignment. Our results in Table III demonstrated that MSC performs better than the other three methods in clustering the sequence and tertiary structure networks. In this study, the performance of MSC was only compared to that of MCL, HC and *K*-means (three of most popular clustering methods). A more general comparison in the performance of various clustering methods with our characteristic threshold D_t is currently under our investigation for various biological networks.

The characteristic threshold distance, D_t, for cluster boundary of MPs identified in this study can be applied to an unsupervised sparsification clustering of MPs. Compared with the regular sparsification clustering of protein networks in previous studies,[17,18] our method has two advantages; 1, the threshold of sparsification is not arbitrary but uniquely identified from the statistical analysis of the network, and 2, most identified protein clusters have unique biological function and are not entangled with other clusters as seen in previous studies. The details of this sparsification clustering of MPs can be found in Table SI of the supporting information. In Figure 4, we showed a visualization of MP clusters with D_t = 1, generated by Cytoscape using the organic layout. In total, there are 111 clusters and 93 isolated nodes. On the whole, for the unsupervised sparsification clustering of MP chains, the average intra-cluster sequence similarity is 0.85 and the average intercluster similarity is 0.02. Two basic types of cluster structure are observed in Figure 4, including a globular structure and an extended structure. To distinguish these two structure types, we defined the cluster connectivity as the number of displayed edges divided by the number of node pairs. A cluster is considered to be globular if its connectivity is greater than 0.85, and extended if its connectivity is less than 0.70. For those clusters whose members are highly similar to each other (for example, FAB, GPCR, potassium/sodium channel, porin, and bacterial rhodopsin), they tend to form a globular structure; and for clusters whose members are only similar to a few neighbors (for example, cytochrome *C* and cytochrome *C* oxidase), they tend to form an extended structure. Some clusters have a mixed structure of both types. For clusters consisting of at least 6 nodes, the average intra-cluster sequence similarity is 0.88 ± 0.08 for globular structures, and is 0.55 ± 0.08 for extended structures. For the largest ten clusters predicted by the sparsification clustering, globular clusters (I, II, III, IV, V, VIII, IX, and X) are consistent with predictions from the second level MSC, as shown in Figure 3.

In Supporting Information Table SIV, we examined the sequence–structure–function relationship for the largest 5 clusters from the unsupervised sparsification clustering method. The largest cluster is composed of 43 antibody Fab fragments, and its average intracluster sequence similarity is 0.97. Cluster II consists of 24 GPCRs, and the average intracluster sequence similarity is 0.91. Cluster III consists of 19 potassium channels and the average intracluster similarity is 0.78. Cluster IV contains 18 ATP-binding cassettes of ABC transporters, and the average intracluster similarity is 0.96. All these four clusters have 100% functional consistency and high

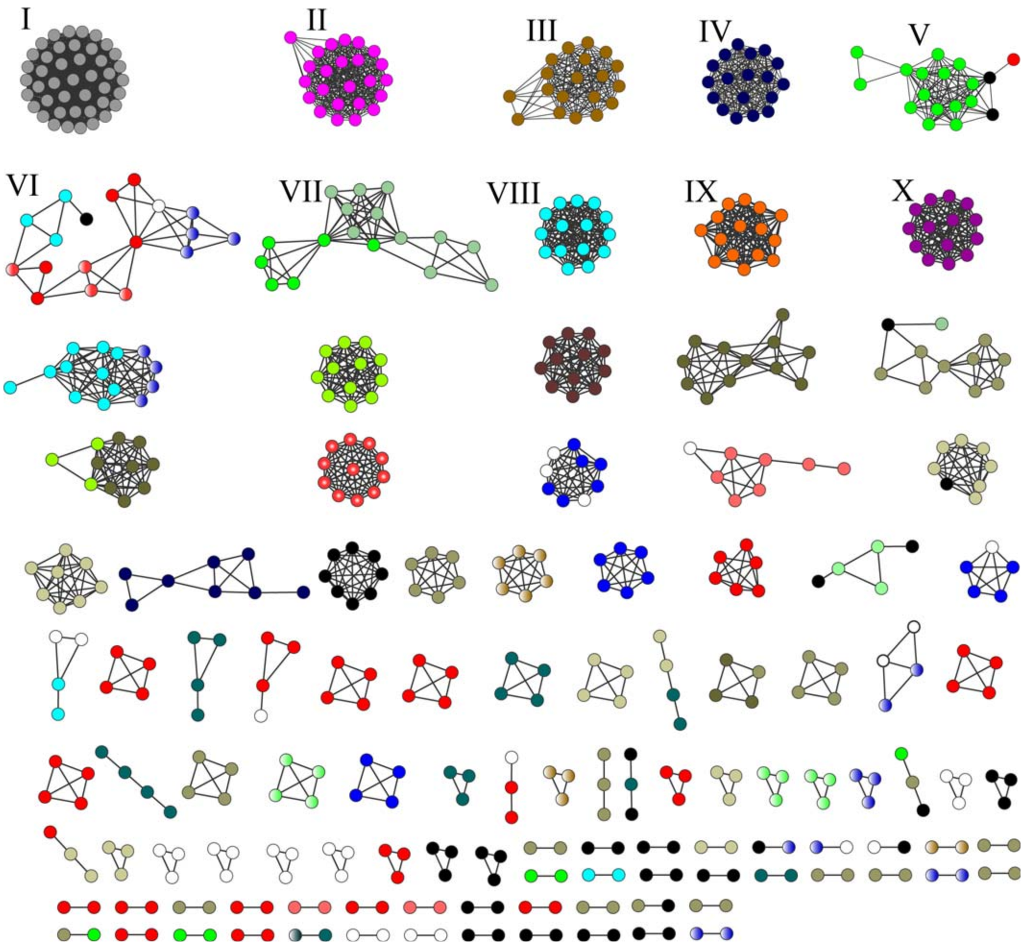

Figure 4

Cytoscape visualization of 111 clusters in the SSN of MPs generated by using the unsupervised sparsification of its distance matrix with D_t = 1. For the largest ten clusters, globular clusters (I, II, III, IV, V, VIII, IX, and X) are consistent with the second level MSC groups, as shown in Figure 4. Only edges of distance less than D_t are displayed. The color code of nodes is the same as that in Figure 2. Isolated nodes are not shown. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

intracluster sequence similarity. Their structures are mainly globular. Cluster V has 18 members and only 15 of them are cytochrome C oxidases. Unlike the other four clusters, this cluster has a mixed globular/extended structure, and thus a lower intracluster sequence similarity and consistency. The above clustering results validate the predicted characteristic threshold distance D_t for the SSN of MPs.

To further study the relationship between sequence and tertiary structure of MPs, in Figure 5, we displayed the tertiary structure similarity matrix of MPs in which network nodes were ordered according to the 1st level MSC sequence clustering (A) or the unsupervised sparsification sequence clustering (B). On these depictions, each matrix element is the tertiary structure similarity (TM score) between two protein chains with a grayscale bar shown on the top. For properly rearranged tertiary structure similarity matrices, the similarity is high in regions near the diagonal, and is low otherwise. The apparent block diagonalization patterns in Figure 5 suggest a strong correlation between sequence and tertiary structure for MPs. Observed off-diagonal signals in Figure 5 indicate similar tertiary structures between proteins in related clusters. For example, in Figure 5(A), the largest cluster and clusters 9, 149, 153, and 156 seem to have similar structures. Chains in these five clusters are all FAB and these clusters merge together in 2nd level MSC and in the unsupervised sparsification clustering. In addition, clusters 2, 7, 69, 71, and 141 also have similar structures. Clusters 2, 69, and 141 are GPCRs and clusters 7 and 71 are bacterial rhodopsins. Although GPCRs and bacterial rhodopsins are not homologous in sequence, they both have 7-transmembrane helix tertiary structures. To quantify similarity of sets of clustering, we calculated their similarity measure as defined in the "Methods" section. The similarity measure is 0.71 between 1st level MSC sequence clustering and sparsification clustering, 0.64 between 1st level MSC sequence

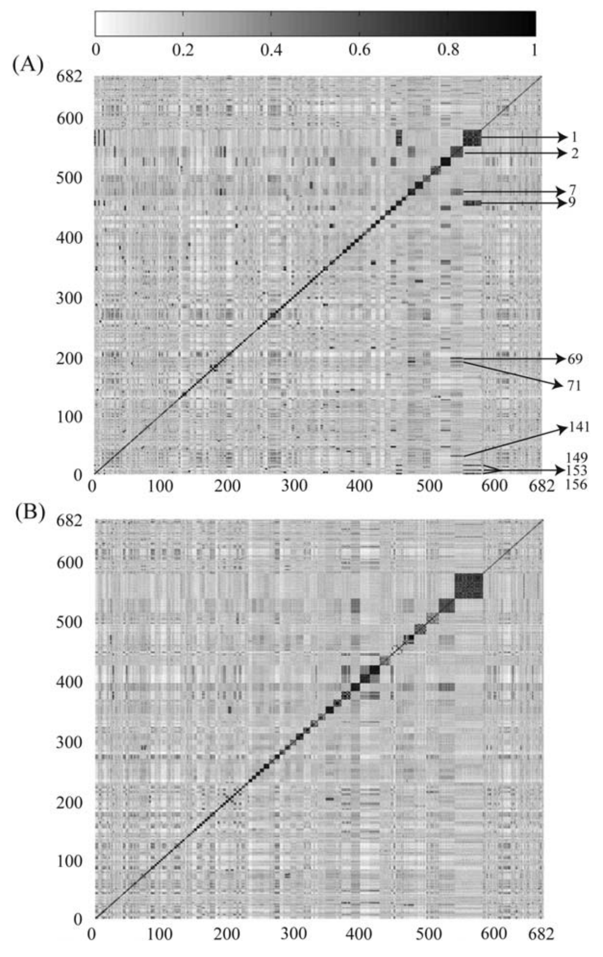

Figure 5

Tertiary structure similarity matrices of MPs, ordered according to their sequence clustering using MSC (A) or the unsupervised sparsification (B). Bar shows the level of similarity in tertiary structure between two MPs with a grayscale from white (0) to black (1). Note that the similarity is very low in the diagonal block for TM chains at the upper right corner, which are isolated nodes in our clustering results.

clustering and 1st level MSC tertiary structure clustering, and 0.50 between sparsification clustering and 1st level MSC tertiary structure clustering. Note that the similarity measure is only 0.32 between two random assignments of 682 nodes into 156 clusters and 111 clusters.

There exist many applications of this study. It is convenient to detect confusing annotations in existing databases. For example, 3m73-A is a voltage-dependent anion channel (VDAC) in the Pfam description and uncharacterized in the GO molecular function, while 3emn-X is a eukaryotic porin in the Pfam description and has VDAC activity in the GO molecular function. Our results showed these two chains have a long sequence distance (7.7) and a low tertiary structure similarity (TM score = 0.27), and belong to different MSC clusters. In the literature, 3m73-A is a SLAC1 anion channel that contains ten transmembrane helices and is weakly voltage-dependent, and differs radically from VDACs (such as 3emn-X) from mitochondrial outer membranes that have a porin-like β-barrel structure.[42,43] Another application is the annotation of previously uncharacterized protein chains, as denoted in Supporting Information Table SI in red color. For instance, in Figure 4, the protein chain 1jb0-K in cluster 29 is uncharacterized in PDB. Since this cluster has a globular structure, it is very likely that 1jb0-K shares the same function as a CBP. To examine this prediction, we compared this protein chain with 37643 structures in PDB and the most similar structure found was 2wsc-G, a CBP in the photosystem I reaction center, with a Z score 4.0 and TM score 0.98. All other structures have a Z score less than 2.0.[44] Furthermore, we conducted a test on the functional identification for tested protein chains based on the sequence information of hub proteins. As shown in Table IV, seven additional protein chains were randomly selected from the UniProt database. For the first six UniProt IDs, their function can be immediately identified by their sequence similarity with hub proteins. For the UniProt ID P04840, its GO molecular functions include "Porin" (UniProtKB-KW annotation) and "Voltage-gated anion channel" (inferred from direct assay[45,46]). Our calculations showed that its sequence distance is 19.54 to the chain 3szd-A (identified as a Porin), and is 0.06 to the chain 3emn-X (identified as VDAC). VDACs are eukaryotic porins located on the

Table IV

Functional Identification of Test Protein Chains by Their Sequence Similarity to Hub Proteins

UniProt ID	Molecular function	Hub ID	Hub function	*E* value	$d_{i,j}$
A2NYU9	FAB heavy chain	3pjs-B	FAB heavy chain	1.4E−23	0.048
P35412	GPCR	4dkl-A	GPCR	6.2E−03	0.046
Q10185	ABC transporter	3d31-A	ABC transporter	2.5E−10	0.032
O14520	Aquaporin	3ne2-A	Aquaporin	3.5E−09	0.016
Q6I9B6	Potassium channel	3ldc-A	Potassium channel	8.9E−04	0.042
Q18DH8	Bacteriorhodopsin	1uaz-A	Bacterial rhodopsins	5.7E−58	0.033
P04840	Porin	3szd-A	Porin	3.0E + 01	19.538
	Voltage-gated anion channel	3emn-X	Voltage-dependent anion-selective channel	6.5E−25	0.056

outer mitochondrial membrane, but their sequences are rather different from sequences of porins located on the outer membrane of gram-negative and -positive bacteria. For porins in the MSC cluster 3, their average sequence distance to 3emn-X and other 43 VDACs randomly selected from Uniport is greater than 90, suggesting a possible convergent evolution between VDACs and bacterial porins. Our annotation of this protein chain as a VDAC is consistent with its UniProt classification based on the direct assay.

The above examples are only for testing the validity of our automated clustering methods for well investigated MPs, and to reproduce known results that are consistent with the literature. Our clustering methods could be particularly useful if the data set consisted of a single superfamily in which all members are evolutionarily related. Using these methods to construct a series of nested networks at different distance thresholds could be useful, because the appropriate distance threshold for separating clusters into groups of proteins that have the same function will be different for different superfamilies.

CONCLUSIONS

The complexity of biological networks surges as enormous experimental data are collected. These biological data are intrinsically variable, noisy, and sometimes imprecise. It is desirable to have improved techniques for the integration and analyses of data arising from different sources, as well as for visualization to understand a wide range of complex networks. In biological networks, this can help identify similar biological entities, like proteins that are homologous in different organisms or that belong to the same complex and genes that are co-expressed. Our proposed clustering methods in this study demonstrated an efficient and convenient approach to identify a general sequence pattern for a group of structure/function related proteins, which could be useful in deriving structural and functional information for novel protein sequences. These methods also offer a panoramic view of protein similarity networks, and allow researchers to trace the relationship between proteins within the same cluster or in neighboring clusters. They could be used as an automated means for classifying proteome or genome databases, which is under our current investigation.

ACKNOWLEDGMENT

CMC thanks P. Pavlidis for the stimulating discussions and hospitality at the Center for High-Throughput Biology, University of British Columbia.

REFERENCES

1. Albert R, Barabási A-L. Statistical mechanics of complex networks. Rev Modern Phys 2002;74:47–97.
2. Watts DJ, Strogatz SH. Collective dynamics of ’small-world’ networks. Nature 1998;393:440–442.
3. Börner K, Chen C, Boyack KW. Visualizing knowledge domains. Ann Rev Inform Sci Technol 2003; 37:179–255.
4. Chang YF, Chen CM. Classification and visualization of the social science network by the minimum span clustering method. J Am Soc Inform Sci Technol 2011;62:2404–2413.
5. Apeltsin L, Morris JH, Babbitt PC, Ferrin TE. Improving the quality of protein similarity network clustering algorithms using the network edge weight distribution. Bioinformatics 2011;27:326–333.
6. Frey BJ, Dueck D. Clustering by passing messages between data points. Science 2007;315:972–976.
7. Samoylenko I, Chao TC, Liu WC, Chen CM. Visualizing the scientific world and its evolution. J Am Soc Inform Sci Technol 2006;57: 1461–1469.
8. Camoglu O, Can T, Singh AK. Integrating multi-attribute similarity networks for robust representation of the protein space. Bioinformatics 2006; 22:1585–1592.
9. Noble WS, Kuang R, Leslie C, Weston J. Identifying remote protein homologs by network propagation. FEBS J 2005;272:5119–5128.
10. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. Nature 2000;407:651–654.
11. Hakes L, Pinney JW, Robertson DL, Lovell SC. Protein–protein interaction networks and biology[mdash]what’s the connection? Nat Biotechnol 2008;26:69–72.
12. MacNeil LT, Walhout AJM. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. Genome Res 2011;21:645–657.
13. Grabmeier J, Rudolph A. Techniques of cluster algorithms in data mining. Data Min Knowl Disc 2002;6:303–360.
14. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv 1999;31:264–323.
15. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. New York: Wiley; 1990.
16. Hartigan J, Wong M. Algorithm as136: a k-means clustering algorithm. Appl Stat 1979;28:100–108.
17. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. Plos One 2009;4:e4345.
18. Uberto R, Moomaw EW. Protein similarity networks reveal relationships among sequence, structure, and function within the cupin superfamily. Plos One 2013;8:e74477.
19. Mashiyama ST, Malabanan MM, Akiva E, Bhosle R, Branch MC, Hillerich B, Jagessar K, Kim J, Patskovsky Y, Seidel RD, Stead M, Toro R, Vetting MW, Almo SC, Armstrong RN, Babbitt PC. Large-scale determination of sequence, structure, and function relationships in cytosolic glutathione transferases across the biosphere. Plos Biol 2014;12:e1001843.
20. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 2002;30: 1575–1584.
21. Vendruscolo M, Dokholyan NV, Paci E, Karplus M. Small-world view of the amino acids that play a key role in protein folding. Phys Rev E 2002;65:6:061910.
22. Atilgan AR, Akan P, Baysal C. Small-world communication of residues and significance for protein dynamics. Biophys J 2004;86:85–91.
23. Krishnan A, Zbilut JP, Tomita M, Giuliani A. Proteins as networks: usefulness of graph theory in protein science. Curr Protein Peptide Sci 2008;9:28–38.
24. Arnold Emerson I, Gothandam KM. Residue centrality in alpha helical polytopic transmembrane protein structures. J Theor Biol 2012; 309:78–87.
25. White SH, Wimley WC. Membrane protein folding and stability: physical principles. Ann Rev Biophys Biomol Struct 1999;28:319–365.
26. Wu HH, Chen CC, Chen CM. Replica exchange Monte-Carlo simulations of helix bundle membrane proteins: rotational parameters of helices. J Comput Aided Mol Des 2012;26:363–374.

27. Membrane Proteins of Known 3D Structure. (2015, April 22). Retrieved from http://blanco.biomol.uci.edu/mpstruc/.
28. Chen CC, Chen CM. A dual-scale approach toward structure prediction of retinal proteins. J Struct Biol 2009;165:37–46.
29. Chen CC, Wei CC, Sun YC, Chen CM. Packing of transmembrane helices in bacteriorhodopsin folding: structure and thermodynamics. J Struct Biol 2008;162:237–247.
30. Huang YH, Chen CM. Statistical analyses and computational prediction of helical kinks in membrane proteins. J Comput Aided Mol Des 2012;26:1171–1185.
31. Mai TL, Chen CM. Computational prediction of kink properties of helices in membrane proteins. J Comput Aided Mol Des 2014;28: 99–109.
32. Kruskal JB. On the shortest spanning subtree of a graph and the traveling salesman problem. Proc Am Math Soc 1956;7:48–50.
33. Wang G, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
34. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
35. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
36. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5?. Bioinformatics 2010;26:889–895.
37. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 2011; 27:431–432.
38. Goldberg DS, Roth FP. Assessing experimentally derived interactions in a small world. Proc Natl Acad Sci USA 2003;100:4372–4376.
39. Torres GJ, Basnet RB, Sung AH, Mukkamala S, Ribeiro BM. A similarity measure for clustering and its applications. Int J Electr Comput Syst Eng 2009;3:164.
40. Pelé J, Abdi H, Moreau M, Thybert D, Chabbert M. Multidimensional scaling reveals the main evolutionary pathways of class a G-protein-coupled receptors. Plos One 2011;6:e19094.
41. Archibald JM, Keeling PJ. Recycled plastids: a 'green movement' in eukaryotic evolution. Trends Genet 18:577–584.
42. Chen Y-h, Hu L, Punta M, Bruni R, Hillerich B, Kloss B, Rost B, Love J, Siegelbaum SA, Hendrickson WA. Homologue structure of the slac1 anion channel for closing stomata in leaves. Nature 2010; 467:1074–1080.
43. Geiger D, Scherzer S, Mumm P, Stange A, Marten I, Bauer H, Ache P, Matschi S, Liese A, Al-Rasheid KAS, Romeis T, Hedrich R. Activity of guard cell anion channel slac1 is controlled by drought-stress signaling kinase-phosphatase pair. Proc Natl Acad Sci USA 2009; 106:21425–21430.
44. Konc J, Cesnik TT, Konc J, Penca M, Janezic D. ProBiS-database: precalculated binding site similarities and local pairwise alignments of PDB structures. J Chem Inform Model 2012;52:604–612.
45. Koppel DA, Kinnally KW, Masters P, Forte M, Blachly-Dyson E, Mannella CA. Bacterial expression and characterization of the mitochondrial outer membrane channel—effects of N-terminal modifications. J Biol Chem 1998;273:13794–13800.
46. Lee AC, Xu X, Blachly-Dyson E, Forte M, Colombini M. The role of yeast VDAC genes on the permeability of the mitochondrial outer membrane. J Membr Biol 1998;161:173–181.