PROTEINS: Structure, Function, and Bioinformatics 61:152–163 (2005)

# A Graph Spectral Analysis of the Structural Similarity Network of Protein Chains

**O. Krishnadev, K.V. Brinda, and Saraswathi Vishveshwara***
*Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India*

***ABSTRACT*** **We present a simple method for the analysis of large networks based on their graph spectral properties. One of the advantages of this method is that it uses a single numerical computation to identify subclusters in a connected graph, which can significantly simplify the complexity involved in analyzing large graphs. This is illustrated using a network of protein chains constructed on the basis of their structural similarities. The large-scale network properties and the cluster and subcluster organization of the protein chain network are presented. We summarize the results of structural and functional analyses of the nodes present in these clusters and elucidate the implications of structural similarity in the protein chain universe. Proteins 2005;61:152–163.** © 2005 Wiley-Liss, Inc.

**Key words: protein chain universe graph; real-world networks; eigenvalue and eigenvector components of graphs; sub-graph density; weighted graphs**

## INTRODUCTION

There has been a great deal of recent interest in understanding the geometry and topology of biological networks.[1] These studies have been guided by the insights gained from the analyses of other real-world networks such as the World Wide Web and social networks.[2] These networks have been described as small-world networks and scale-free networks based on some of their simple characteristics. For example, scale-free networks exhibit a power-law degree distribution rather than the exponential distribution expected for random networks.[2] A similar analysis has been carried out on biological networks including metabolic networks,[3] transcription regulatory networks,[4] protein–protein interaction networks[5] and protein domain networks.[6,7] Most of these networks may be classified as scale-free networks. One of the common features of all these networks (biological and nonbiological) is that they are very large with thousands of nodes and edges and hence their analysis becomes very intensive and difficult. With the growing complexity of biological networks due to the availability of many genome sequences, simpler methods of analyzing such large networks are required. We present one such simple method in this paper, which uses a combination of traditional clustering algorithms like the depth first search (DFS) and a graph spectral method to identify clusters and subclusters progressively in a large network. Such an algorithm helps in understanding the large-scale properties of the networks as well as in analyzing the details of the connections seen in the networks at the individual node level to understand their biological implications. Hence, the methodology presented in this paper could prove to be useful to systems biologists and genome biologists. The graph spectral method presented here has already been used to identify amino acid clusters of biological significance in protein structures[8,9] and also to automatically partition multidomain protein structures into individual domains.[10] The identification of structural domains in a multidomain protein uses the fact that the amino acids contacts are much higher within the domain than across domains. Hence, in a connected graph representation of protein structure, the spectra of the graph gives information regarding the subclusters in the graph, which in the case of multidomain proteins happen to be the domains. Hence, the nodes forming the subclusters in the graph are partitioned into structural domains.

We illustrate the power of the graph spectral method by analyzing the protein chain universe graph (PCUG), constructed based on the structural similarities between protein chains. Earlier, a similar network was constructed and analyzed by Dokholyan et al.[6] However, they used protein domains rather than protein chains in their analysis in order to obtain an evolutionary perspective on the protein domains. The motivation for the use of complete protein chains rather than protein domains comes from the fact that more than 80% of the eukaryotic genomes are composed of multidomain proteins,[11] where each protein is made up of more than one domain contributing to the same overall biological function. Hence, the analysis of protein structures beyond the domain level is required to understand the organization of protein chains in structure space and the combinations of protein domains in the chain universe. Keeping this in mind, we have generated the PCUG using the DALI Z score[12] as the structure similarity index and have analyzed some of the basic network properties of the PCUG and the biological properties observed in the clusters and subclusters of PCUG. Some of

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

*Correspondence to: Prof. Saraswathi Vishveshwara, Molecular Biophysics Unit, Indian Institute of Science, Bangalore, 560012 India. E-mail: sv@mbu.iisc.ernet.in

Received 27 September 2004; Revised 28 January 2005; Accepted 28 January 2005

Published online 3 August 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20532

the specific clusters obtained in the PCUG are presented in detail to elucidate the advantages of the graph-spectral method and to obtain some biological insights on the chain universe.

## MATERIALS AND METHODS

### Data Set

The data set for the present work was obtained from the FSSP representative set, release Dec 2004.[13] It contains protein chains present in PDB90[14] (PDB with sequence identity cutoff of 90%). The dataset was clustered using the program blastclust[15] to obtain a nonredundant set at 25% sequence identity. For each protein chain in the representative set, FSSP gives all its structural neighbors (defined as chains showing DALI Z score greater than 2) and their respective DALI Z scores. The greater the DALI Z score between a pair of chains, the more similar are the structures of the two chains.

### Construction of PCUG

The Protein Chain Universe Graph (PCUG) is constructed as follows. Each chain is a node and those nodes with a structural similarity score above a chosen Z cutoff (called $Z_{min}$) are connected by an edge. All edges/connections are given equal weight of one. In order to perform numerical analysis on the PCUG, it is represented as an adjacency matrix, which is an $n \times n$ matrix in which $n$ is the no of nodes (chains) in the graph. The adjacency matrix $[A]$, is created with the following rules:

$$[A]_{ij} = 1 \text{ if } Z_{ij} \geq Z_{min}$$

$$= 0 \text{ if } Z_{ij} < Z_{min}$$

$$= 0 \text{ if } i = j$$

where $[A]_{ij}$ is the $ij^{th}$ element of the matrix and $Z_{ij}$ is the DALI Z score between chains $i$ and $j$. For generation of weighted matrices, the value of $[A]_{ij}$ is set to the Z score of the domain pair. The adjacency matrix is symmetric because the structural similarity is a symmetric relationship. PCUGs were generated using different $Z_{min}$ values varying from 2 to 20 and were analyzed using DFS and graph spectral methods as explained in the next few subsections.

### Clustering of PCUG Using DFS

DFS (Depth First Search) graph algorithm[16] is one of the traditionally used methods to identify clusters in a graph. This method is used here to obtain information on the clusters obtained in the PCUGs generated at $Z_{min}$ values 2 to 20. DFS identifies the disjoint clusters present in PCUG. The clusters are then visualized using GraphViz[17] to obtain the details of the connectivity within the cluster. The clusters were then individually analyzed using the Graph Spectral method, as described in the next section.

### Graph Spectral Analysis

Graph spectral analysis is a sub-branch of graph theory dealing with the analysis of the spectra (eigenvalues and eigenvector components) of the nodes in the graph. Such an analysis can reveal information about the global arrangement of nodes in the graph and hence can be very useful in the analysis of large graphs. A brief description of the analysis is given here. To generate the eigenvalue spectra of the graph, the adjacency matrix of the graph (as given in the previous section) is converted to a Laplacian matrix, which is defined as the matrix obtained by subtracting the adjacency matrix from the degree matrix (L = D − A). The degree matrix is a diagonal matrix in which the $i^{th}$ element on the diagonal is equal to the no of connections, which the $i^{th}$ node makes in the graph with other nodes. For the weighted degree matrix, the diagonal element is equal to the sum of weights of all connections emerging from the node.

Diagonalization of the Laplacian matrix yields the spectra of the graph comprising of the eigenvalues and the corresponding eigenvector components. The analysis of the vector components of the lower and the higher eigenvalues have been shown to give information about the clustering of nodes in the graph and the connectivity of each node.[8,9] The second lowest eigenvalue and the vector components corresponding to it yield information about the clusters present in the graph with all nodes of a given cluster having the *same* value of the vector component.[8,9] Interestingly, in a completely connected graph, where all nodes belong to a single cluster, the vector components of the second lowest eigenvalue give subcluster information, where the nodes within a subcluster have *similar* vector component values. A "subcluster" is defined as a set of nodes within the cluster, which make significantly more connections among themselves than with the other nodes in the cluster. A plot of the sorted vector components can bring out the subcluster information very well. In this plot, the nodes that are part of a subcluster show up as distinct plateaus on a curve that otherwise show a monotonously increasing behavior.

### Structural and Functional Analysis

The structural classification of the proteins present in the clusters and subclusters obtained using DFS and the graph spectral method is identified from the SCOP database.[18] The functions of these proteins are identified from the FSSP file of the chain, which reports the COMPND record from the corresponding PDB file. This is useful for the analysis of the structure–function correlations between the chains present in a cluster.

## RESULTS AND DISCUSSION

The protein chain universe graph (PCUG) is constructed from a nonredundant set of protein chains (3477) with known structures, using the DALI Z score as the edge-forming criterion (given in detail in Method section). The clusters in the PCUG obtained for different $Z_{min}$ scores have been identified using the DFS method. The details of the protein chains present in the clusters obtained at $Z_{min}$ score 11 are given in a supplementary table. The properties of the PCUG in terms of scale-free and random behavior are presented in the next section. Further analysis of the intricate details and the nature of connectivity is

**TABLE I. Cluster Sizes Obtained From PCUGs at Different $Z_{min}$ Values†**

| $Z_{min}$ value | Number of orphans | Sizes of the top five large clusters |
|---|---|---|
| 2 | 182 | 3178,23,12,10,7 |
| 3 | 328 | 2946,19,12,9,7 |
| 4 | 520 | 2607,20,14,11,10 |
| 5 | 721 | 2263,14,12,11,10 |
| 6 | 944 | 1932,20,13,12,12 |
| 7 | 1150 | 1559,16,15,12,12 |
| 8 | 1318 | 1134,37,37,29,25 |
| 9 | 1463 | 916,36,25,23,20 |
| 10 | 1620 | 565,48,36,26,23 |
| 11 | 1747 | 326,45,40,35,30 |
| 12 | 1873 | 289,34,34,28,25 |
| 13 | 2010 | 147,92,23,22,15 |
| 14 | 2115 | 122,79,23,20,15 |
| 15 | 2224 | 81,75,27,23,16 |

†From DFS method.

presented below in the section "Subcluster identification form the second-lowest eigenvalue." Finally, an analysis of the structural and functional correlations embedded in the clusters and subclusters of PCUG is presented below in the section "Structural and Functional Features of the Clusters Seen in PCUG."

## Network Properties of PCUG

The details of clusters in PCUG obtained at $Z_{min}$ scores of 2 to 15 are given in Table I. As expected, the number of orphans (nodes with zero connections) increases and the graph becomes more disjoint as $Z_{min}$ increases. The network properties of PCUG are analyzed in terms of the parameters presented below.

### *Size of the largest cluster and degree distribution as functions of $Z_{min}$*

The size of the largest cluster is plotted as a function of $Z_{min}$ in Figure 1(A), which shows a transition around the $Z_{min}$ of 11. We have also analyzed the degree distribution of the nodes in the PCUGs generated at various $Z_{min}$ values. We find that at lower $Z_{mins}$, PCUG shows the behavior characteristic of a random graph with the degree-distribution not following power-law behavior. However, as $Z_{min}$ is increased, the number of connections in the graph decreases and approximate power-law behavior is observed in the log–log plots of the degree distribution at higher $Z_{min}$ values (beyond 6). However, the best fit to the power-law is seen at $Z_{min} = 11$ with an exponent value of 1.8. The log–log plot of the degree distribution in PCUG at $Z_{min} = 11$ is shown in Figure 1(B), where the curve is clearly linear. The protein domain universe graph (where each protein domain was considered as a node) analyzed by Dokholyan et al.,[6] was also found to show a power-law degree distribution at a cutoff $Z_{min}$ of 9, around which it is scale-free. Further, the power-law fit deteriorated both above and below the $Z_{min}$ of 9. However, the number of orphans remains dominant in case of the domain graph at all $Z_{min}$ values. The protein chain universe graph presented here shows power-law degree distributions only at higher $Z_{mins}$ and orphans do not dominate at lower $Z_{mins}$. Hence the PCUG is random at lower $Z_{mins}$ and becomes scale-free only at higher $Z_{mins}$. This is similar to the behavior observed in structure space of the lattice model,[19] where the scale-free behavior is seen only above a threshold value of the similarity score.

### *Degree density*

The degree density of a connected graph is evaluated as the ratio of the average number of connections observed per node and the maximum number of connections possible if the graph was a clique (a complete graph where each node is connected to every other node). The maximum degree density value is one, which is seen in the case of cliques. The degree density of the PCUG is evaluated at different $Z_{mins}$ and the plot of the degree density versus $Z_{min}$ is shown in Figure 1(C). The log–log plot of the same figure is shown in Figure 1(D). Figure 1(C,D) shows that the degree density versus $Z_{min}$ also follows power-law behavior with an exponent of 2.2.

We alert the reader that the results presented here pertain to the limited dataset currently available, which has many truncated proteins. The proportion of multidomain chains in PCUG is around 30% whereas in the actual chain universe, it is somewhere around 65%.[11] This is due to the fact that multidomain proteins are not easy to study by existing methods of structure determination. For example, the structure of almost all the domains of Protein Kinase C has been solved, but the structure of the full chain is still not available. Hence the PCUG generated here is not a complete representation of the actual chain universe and our evaluation of this network is subject to the constraint of data limitation. However, for the actual chain universe with many more multidomain chains, one would necessarily have a larger number of connections and the connections that we see here are a subset of the actual chain universe. Hence, the structural similarities seen here are genuine and will not be affected due to the presence of truncated proteins.

Henceforth, we will present analysis of PCUG at $Z_{min} = 11$ because the size of the largest cluster Vs. $Z_{min}$ shows a transition at $Z_{min} = 11$ and the power-law fit to the degree distribution is best at this value. Moreover, we get clusters with sizes that can be easily handled for individual node and connectivity analyses at this $Z_{min}$.

## Subcluster Identification from the Second-Lowest Eigenvalue

Clusters in a network can be differentiated into two types based on the organization of connections between the nodes present in them. We define a "simple cluster" as one in which the nodes do not form any distinguishable subclusters and a "complex cluster" as one in which there are two or more distinguishable subclusters. Many of the large-scale networks have both simple and complex clusters and hence the identification of subclusters makes the analysis of the large networks easier. The graph spectral algorithm presented here (as explained in the Materials

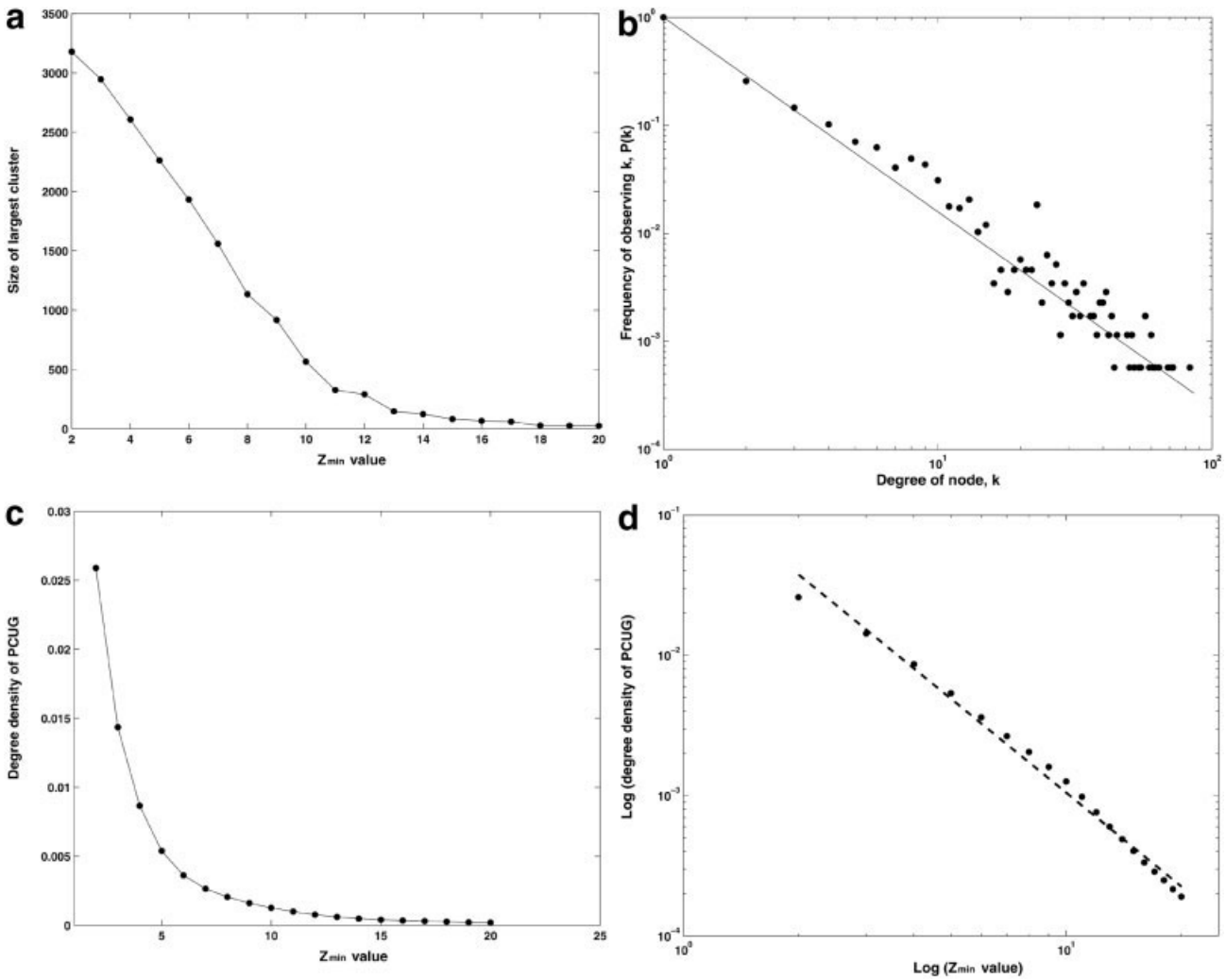


Fig. 1. **A**: A plot of the size of the largest cluster obtained in PCUG as a function of $Z_{min}$ values. **B**: Log–log plot of the normalized degree distribution function *P(k)* versus the degree, *k* at $Z_{min} = 11$. The values observed are normalized in such a way that the probability of observing nodes with zero connections is 1. The figure also shows a fitted line with equation, $P(k) = k^{-1.8}$. **C**: A plot of the degree density as a function of $Z_{min}$. Degree density is defined as the ratio of the average number of connections per node and the maximum number of connections possible per node in the graph. **D**: Log–log plot of the degree density versus $Z_{min}$ (see C). The fit to the curve (dashed line) shows power-law behavior and has an exponent of 2.2.

and Methods section) gives an objective way to obtain the subclusters in a connected cluster, where the vector components of the second lowest eigenvalue (represented by 2evc henceforth) give information about the subclusters. The identification of subclusters from the eigenvalue spectra can be done easily by plotting the sorted eigenvector components of the second lowest eigenvalue versus node number (the 2evc plot). The nodes forming subclusters show up on this plot as having the same or very similar vector component magnitude. As an example, Figure 2(A) shows the vector component plot of cluster 11 (all examples and the cluster numbers are taken from the supplementary table) showing two distinct regions in the plot. The graph layout, generated using the software GraphViz,[16] is given in Figure 2(B), which confirms the information provided by the vector component plot. It can be seen that the nodes, which have similar vector component values in the plot form a subcluster, which is clearly distinct for the other subcluster.

The identification of subclusters within a cluster in PCUG is of importance because of two reasons. First, the PCUG is a large graph and hence obtaining subcluster information of the bigger clusters makes the analysis easier as presented in this section. Second, even in the smaller clusters, the partitioning of subclusters could give biologically relevant information regarding the structure–function correlations between the proteins belonging to the subclusters as discussed in a later section. The graph spectra of a few representative clusters are presented here to elucidate the application of this method in the identification of subclusters in a connected cluster.

### *Simple clusters*

The simple clusters can be classified on the basis of their degree density as those with high or low degree density. Cluster 13 having 18 nodes is an example of a simple cluster with low degree density (0.2) lacking subclusters.

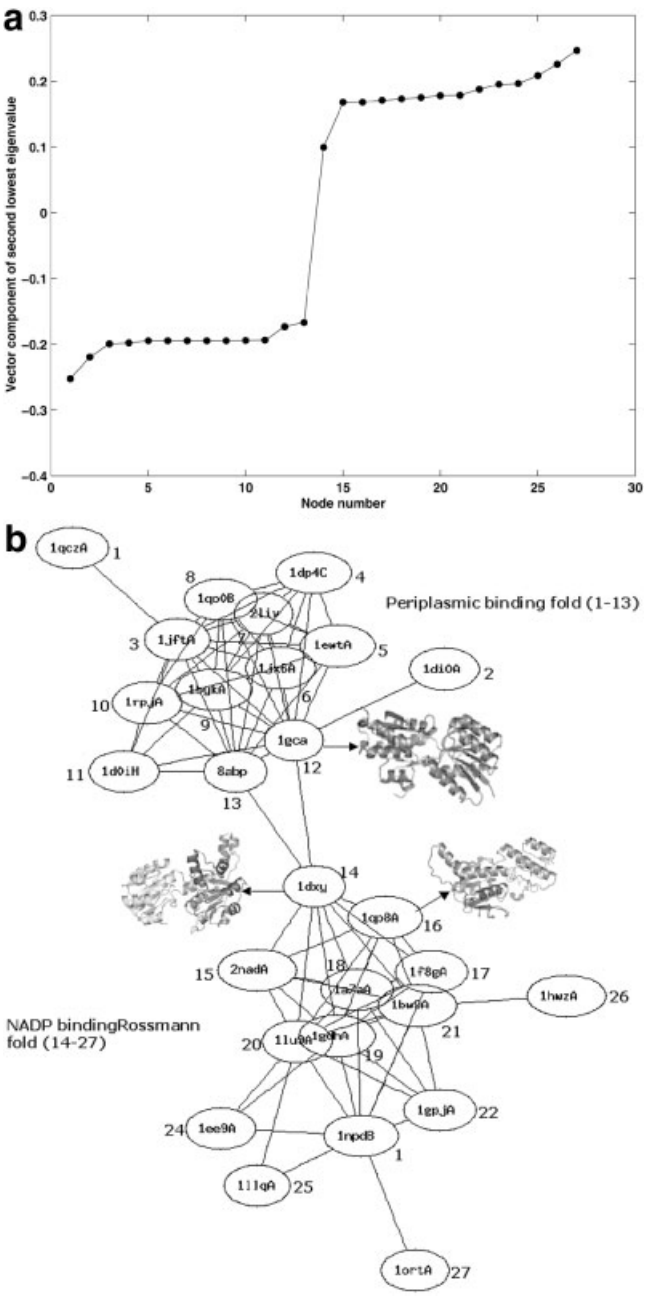


Fig. 2. **A**: A plot of the sorted vector component of second lowest eigenvalue of each node in Cluster 11 obtained at $Z_{min}$ 11. Such a plot is denoted as 2evc plot in subsequent figures for other clusters. The cluster 11 has two distinct subclusters and this is can be identified by the two distinct plateaus in this vector component plot. **B**: Graph layout of Cluster 11 showing the nodes (with PDB codes) of the cluster with node numbers corresponding to the ones given in (A). The folds represented in this cluster are indicated along with the node numbers of those proteins belonging to the fold. Same convention is followed in subsequent graph layout figures.

The 2evc plot of cluster 13 is given in Figure 3(A) and the graph layout is given in Figure 3(B). The graph layout also gives additional information regarding the domains present in the nodes. The graph layout in Figure 3(B), points out that there are not many connections among the nodes in this graph and they do not form distinct subclusters. Thus, it can be deduced that this cluster lacks proper subclusters. The 2evc plot reflects this behavior except for the nodes 15, 16, 17, and 18 for which the vector component values are the same. These nodes are not part of a subcluster but still show up as having the same vector component value. The rest of the nodes show a monotonous increase in the vector component value reflecting the fact that these nodes are "loosely" placed in the graph.

Cluster 24 having 12 nodes is a simple cluster with high degree density (0.88). Usually the clusters with high degree density form a near-clique with densely connected subclusters. The 2evc plot of cluster 24, shown in Figure 4(A), clearly demonstrates that the cluster is a simple one with almost all the nodes forming a plateau in the 2evc plot. This clearly indicates that all the nodes form a densely connected subcluster. This can be visualized from the graph layout given in Figure 4(B).

## Complex clusters

Complex clusters by definition are composed of two or more subclusters with very few connections across the subclusters. Hence, the degree density is generally low in the complex clusters. Cluster 4, which is the largest cluster obtained at $Z_{min}$ value of 11, is a complex cluster having subclusters. The 2evc plot of this cluster and the graph layout of cluster 4 are given in Figure 5(A,B), respectively. The 2evc plot indicates the presence of at least three subclusters in this cluster, which can be identified from the graph layout in Figure 5(B). Analysis of the individual subclusters shows that the subcluster 3 is further made up two subclusters as seen in Figure 5(B). Though this segregation is not so clear in Figure 5(A), we do find a kink in the plateau of subcluster 3, which corresponds to this split into further subclusters. This further subclustering of subcluster 3 is more clearly seen when this subcluster is separately analyzed rather than as a part of the whole big cluster. Thus, the 2evc plot is invaluable for the objective identification of the subclusters in a large cluster, which makes further structural and functional annotation simpler.

In this section, we have presented an analysis of simple and complex clusters in order to show that the subcluster information can be obtained from the graph spectra. The subcluster information is validated when the $Z_{min}$ cutoff is increased or decreased. At higher $Z_{min}$ values, the subclusters break off into individual clusters. At lower $Z_{min}$ values, many clusters merge with each other and form subclusters. This can be seen from Figure 6, where two clusters of five and 11 nodes each at $Z_{min}$= 11, are connected into a single cluster at $Z_{min}$ = 8 (with the total number of nodes in the combined cluster being 24 due to the presence of other nodes which are orphans at $Z_{min}$ 11). Further, the 2evc plot of this cluster at $Z_{min}$ = 8 (Fig. 6), gives two subclusters corresponding to the two clusters obtained at $Z_{min}$ = 11. This clearly elucidates that subclustering at lower $Z_{min}$ values can validate the clustering information obtained at higher $Z_{min}$ values since the subclusters present in a single cluster at a lower $Z_{min}$

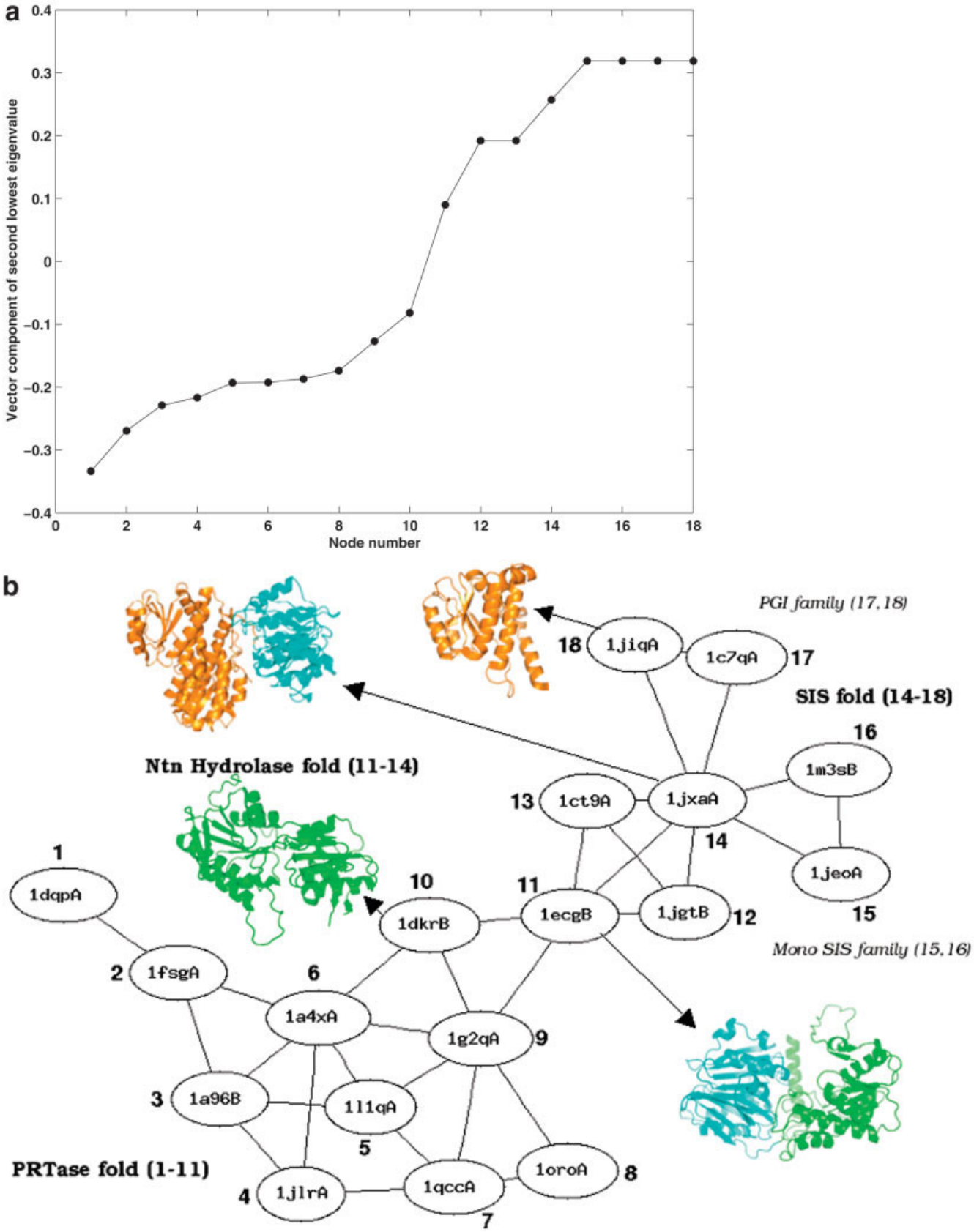


Fig. 3. **A**: 2evc plot of Cluster 13 obtained at $Z_{min}$ 11. The cluster has no distinct subclusters and is of low degree density. **B**: Graph layout of Cluster 13 showing the nodes (with PDB codes) of the cluster with node numbers corresponding to the ones given in (A). The protein structures of some of the nodes are shown in the figure and these are indicated by arrow from the node to the protein structure. The domains are shown in different colors and correspond to the different folds present in these proteins.

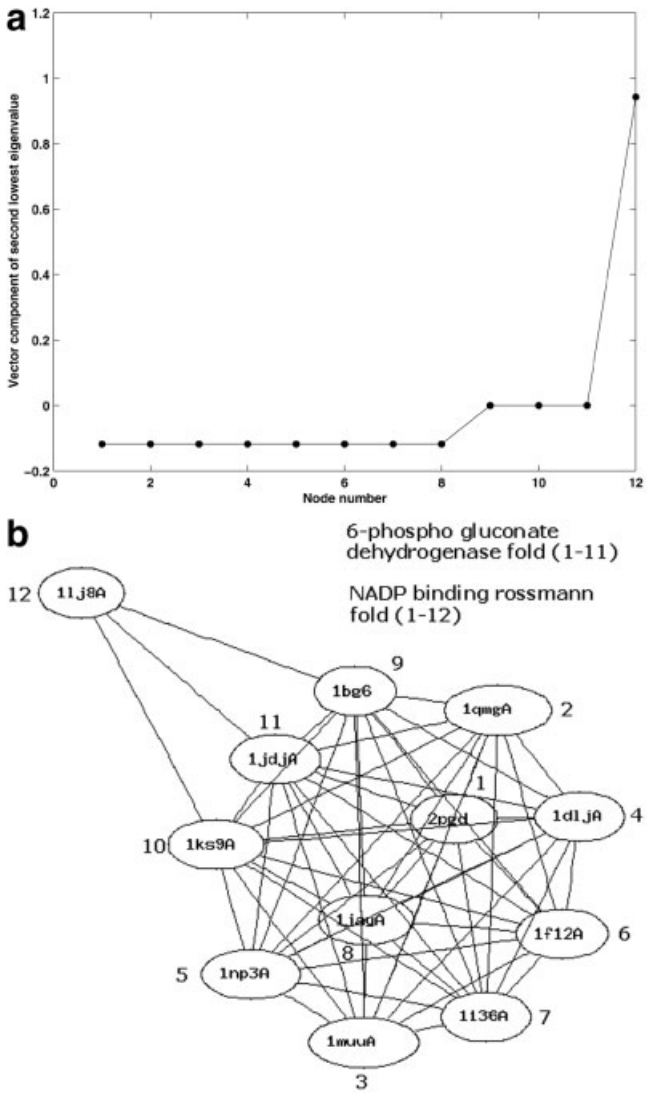


Fig. 4. **A**: 2evc plot of Cluster 24 obtained at $Z_{min}$ 11. The cluster has two closely linked subclusters that are densely connected and this is suggested by the fairly close regions of vector component values for the nodes corresponding to those nodes. **B**: Graph layout of Cluster 24 showing the nodes (with PDB codes) of the cluster with node numbers corresponding to the ones given in (A).

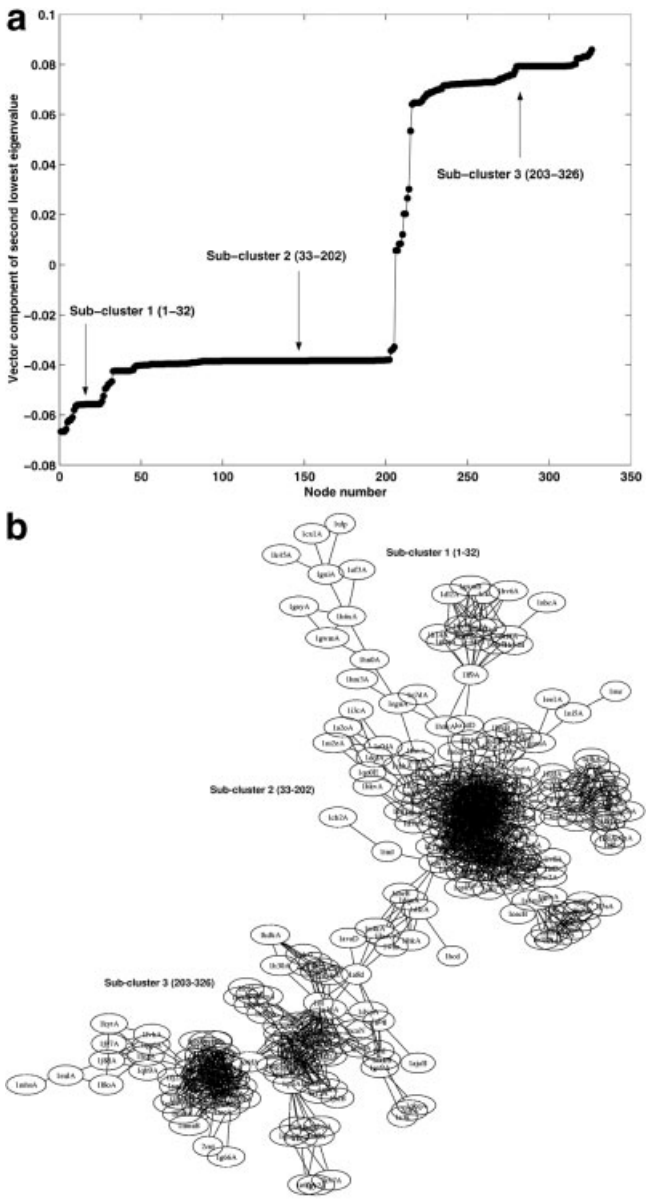


Fig. 5. **A**: 2evc plot of Cluster 4 (largest cluster in PCUG at $Z_{min}$ 11). The cluster has three subclusters and these are indicated on the plot with the numbers of the nodes belonging to that subcluster. **B**: Graph layout of Cluster 4 showing the nodes (with PDB codes) of the cluster. The three subclusters identified in this cluster are indicated in the figure. Due to the large size of the cluster, individual nodes are not marked in the figure.

become separate clusters at a higher $Z_{min}$. Using algorithms like DFS one can get the information about the disjoint clusters present at a particular $Z_{min}$ value but the subclustering information cannot be directly obtained. Repeated clustering of nodes using DFS at various $Z_{mins}$ can give the subcluster information in the PCUG. However, the graph spectral method gives the subcluster at any $Z_{min}$ in a single step. Thus when analyzing large graphs, where the information about the organization of the nodes in a cluster is required, graph spectral analysis scores over other methods.

In principle, the PCUG could be generated at a single low $Z_{min}$ cutoff, and from the graph spectra, one could locate the subclusters, which at higher $Z_{min}$ scores will become individual clusters. However, clustering by the graph spectral method in large graphs suffers from a limitation due to the resolution of the vector component values. Because the normalized value of the sum of the vector components is 1, a larger number of nodes, in general, result in very small values of the vector components. In practice, this can be overcome by initially segregating the distinct clusters in a large graph and then break the large clusters obtained into smaller fragments (subclusters). This paper elucidates one such method, which uses DFS for getting distinct clusters and a graph spectral algorithm for obtaining the subclusters. The smaller fragments obtained using this method are very easy to analyze

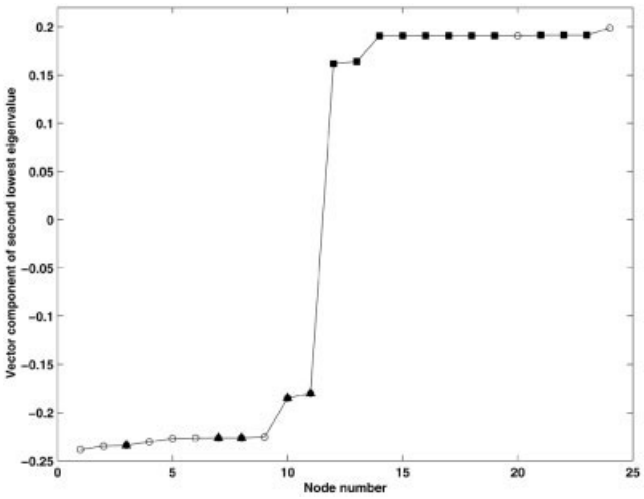


Fig. 6. 2evc plot (solid line) of a cluster with 24 nodes at $Z_{min}$ 8. This cluster splits up into two clusters of sizes 11 (■) and 5 (▲) respectively and eight orphans (○) at $Z_{min}$ 11.

and hence, this method has the potential to be very useful in the analysis of large graphs. Another useful method is the construction and analysis of edge-weighted PCUG with the $Z_{min}$ scores as the edge weights in the graph. This would generate a single PCUG with the $Z_{min}$ scores embedded as edge-weights in the graph. However, we do not discuss these edge-weighted PCUGs in this paper.

### *Analysis of the vector components of the highest eigenvalue*

Apart from the identification of clusters and subclusters, graph spectra can also be used in the identification of cluster centers. The cluster center is defined as the node, which has maximum connection in the graph and is also close to the geometric center of the graph from which the distance to every other node is minimum.[9] The cluster center can be identified by the analysis of the vector components of the highest eigenvalue (referred to as hevc henceforth) and is found to have highest magnitude of hevc.[9] Hence, in the plot of hevc versus node number (the hevc plot), the cluster center appears as a peak. For example, the hevc and 2evc plots of Cluster 7 having 14 nodes is shown in Figure 7(A). The graph layout indicating the node numbers and the PDB codes is given in Figure 7(B). This cluster has no subclusters as is evident from the 2evc plot and the graph layout. It can be seen from Figure 7(A) that Node 7 has the highest eigenvalue vector component in the hevc plot. This node (1opoC) has maximum connections in the cluster as can be seen from Figure 7(B) and the sum of the distances from this node to any other node in the cluster is the least. Thus, it is the cluster center and has been identified in a single computation by graph spectral analysis.

## Structural and Functional Features of the Clusters Seen in PCUG

The graph representation of biological networks and their analysis has gained popularity in recent times due to

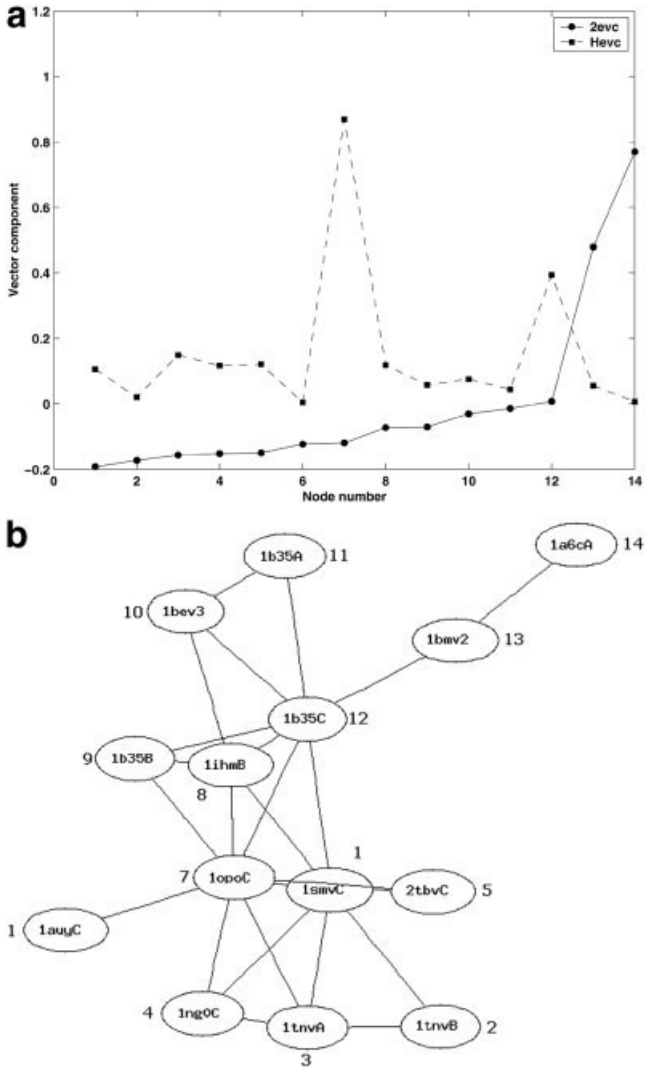


Fig. 7. **A**: Plot of vector component of the highest and the second lowest eigenvalue for each node of Cluster 7 obtained at $Z_{min}$ 11. The 2evc plot indicates that there are no distinct subclusters present in this cluster. The hevc plot showing the vector component of highest eigenvalue for each node is also given in the figure. The hevc plot suggests that node 7, which has the highest magnitude in this plot, is the cluster center. **B**: Graph layout of Cluster 7 showing the nodes (with PDB codes) of the cluster with node numbers corresponding to that given in Figure 6(A).

the fact that they give a view of the large-scale organization of the biological components. Some of the networks occurring in nature cannot be understood by studying individual components alone, as they do not give insights into the organization of networks. Examples include signal transduction cascades, gene regulatory networks and protein interaction networks where a single component cannot be understood without finding the "context" of its function in the network. Another example is the analysis of evolution of proteins, which exist in nature at present. Much work has been done on the evolution and classification of proteins at the domain level. It has been demonstrated that the "Domain Universe" is largely "scale-free."[6,7] The analysis of domains in itself is significant but most proteins in the genomes studied are multidomain

proteins and therefore it is important that protein evolution be studied both at the domain level and at the chain level. Domain recombination and duplication are the processes that drive the formation of new proteins with novel functions. The recombination can either be random or otherwise. The random recombination of domains might lead to counter-productive domain combinations, and hence these combinations will be eliminated by natural selection. Thus, the existing combinations are likely to have arisen due to strong selection processes. However, there are also arguments in favor of random recombination of domains. These aspects related to domain combination events are discussed in the review by Koonin and coworkers.[20] Since the mechanisms that govern domain combinations are not clearly understood, studies related to domain fusion is an interesting area of research, which can enhance our understanding of the structure-function relationships in proteins.

In the analyses presented in this paper, we have found that the connections between the chains forming clusters in PCUG can be due to (1) the presence of domains with the same fold or (2) the presence of multidomain chains consisting of two different folds that bring together different folds into the same cluster or (3) structural similarity between two different folds. A few interesting cases where the nodes have formed clusters or subclusters with biologically interesting properties are explained below. Our primary aim here is to assess the efficiency of the graph spectral method as applied to the PCUG.

### *Structural features of protein chains clustered in PCUG*

DALI Z score is known to distinguish two different folds even at a $Z_{min}$ of 2.[21] Thus, it is interesting to find two different folds being present in the same cluster at a high $Z_{min}$ cutoff of 11. We analyzed such cases further and Cluster 13 [Fig. 3(A,B)] is a good example of a cluster where we find two different folds present in the same cluster at $Z_{min} = 11$. In this cluster, three different SCOP folds are present namely SIS domain, PRTase domain, and Ntn Hydrolase domain. Node no 11 (1ecgB) has both Ntn hydrolase domain and PRTase domains. Nodes 1–10 have only PRTase domains. Nodes 11–14 have Ntn hydrolase domain. Node 14 has both Ntn hydrolase domain and a SIS domain and nodes 15–18 have SIS domains. It is clear from this example that due to the presence of two multidomain chains namely 1ecgB (PRTase and Ntn hydrolase) and 1jxaA (Ntn hydrolase and SIS domain), three different SCOP folds have come together in this cluster. It can be seen from Figure 3(B) that the SIS domain nodes 15–18 are partitioned into different subclusters where nodes 15 and 16 form one subcluster and nodes 17 and 18 form another subcluster. They do not have any direct connections with each other and are connected only through Node 14. On further analysis, it was found that nodes 17 and 18 belong to a different family (Phosphoglucose isomerase family) than nodes 15 and 16 (mono SIS family) though they all belong to same SIS domain fold. It is interesting that the chains belonging to different families of the same fold get partitioned in the clusters seen in PCUG.

The segregation of chains within a cluster according to SCOP family is also seen in Cluster 14, which has 21 nodes. The 2evc plot and the graph layout of this cluster are shown in Figure 8(A,B), respectively. In this cluster, only PH domains and DBL homology domains are found. Nodes 1, 2, and 5 have both PH domain and DBL homology domain. Nodes 3 and 4 have only DBL homology domain. Nodes 6–21 have only PH domains. The chains with PH domains and DBL homology domains have come together due the presence of multidomain chains containing both the folds. Figure 8(A) shows the partitioning of the subclusters according to presence of PH and DBL homology domains. Further analysis of the PH domain subcluster shows the segregation of all the PH domain folds according to their families as can be seen from Figure 8(B). Nodes 6–8 belong to Phosphotyrosine binding family, Nodes 10 and 11 belong to Acyl Coa binding family, node 18 belongs to Ran-binding family, nodes 19–21 belong to the VASP/Enabled homology family and the other PH domain containing chains belong to the Pleckstrin Homology family. This segregation is due to the fact that DALI Z scores are able to distinguish the structural changes occurring across SCOP families belonging to the same fold. The present analysis clearly brings out such interesting features observed in the chain universe.

Subcluster 1 of the biggest cluster (cluster number 4) has 32 nodes comprising of chains with folds belonging to all β class of SCOP. The folds represented the most are immunoglobulin-like β sandwich, β-trefoil and concanavalin A fold. It is interesting to note that Cluster 1 (Supplementary Table) also has chains with immunoglobulin-like β sandwich fold, but the proteins seen in the nodes of Cluster 1 are part of the immune system or signal transduction pathways and are functionally different from the ones seen in this cluster. This clearly suggests a structural segregation of functionally different chains, which have the same SCOP fold. This feature appears because the DALI Z scores are sensitive to the structural differences in such proteins and is reflected in this graph spectral analysis.

Different folds occurring in the same cluster due to multidomain chains is expected in a chain universe graph. However, what is not expected is the presence of two different folds in a cluster although there is no multidomain chain containing the two folds. Cluster 11 [Fig. 2(A,B)] is an example showing such connections. This cluster consists of two subclusters belonging to Periplasmic binding domain and NADP-Rossman fold. The common link between the two subclusters in node 14 (1dxy), which has NADP-Rossman fold and Flavodoxin fold. This node connects to the Periplasmic-binding subcluster through nodes 12 and 13 (8abp and 1gca), both having the Periplasmic binding fold. All the three domains (NADP-binding Rossman fold, Flavodoxin fold and Periplasmic binding fold) belong to SCOP α/β class and have been connected in the cluster due to structural similarities arising due to similar arrangement of the secondary

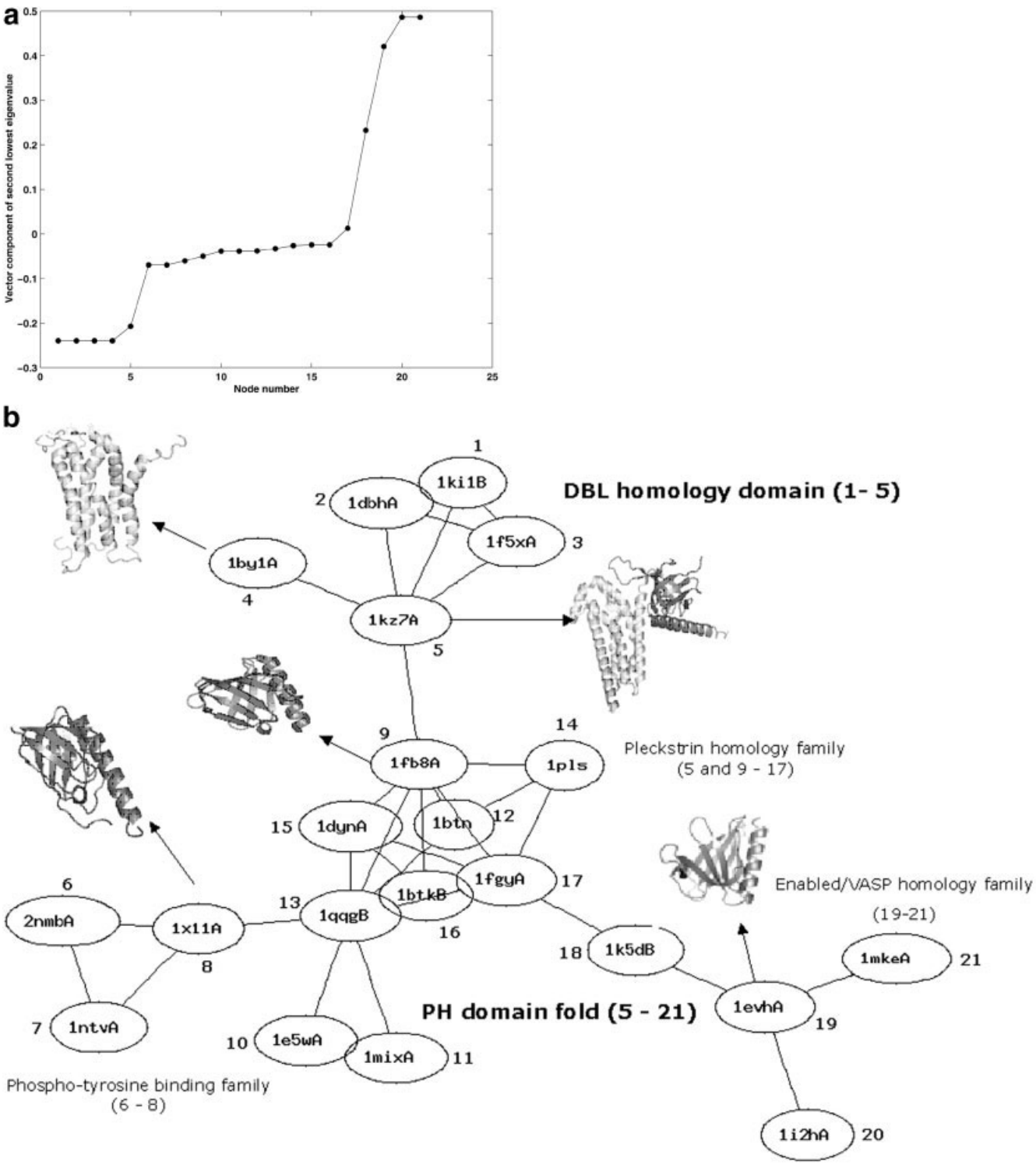


Fig. 8. **A**: 2evc plot of Cluster 14 obtained at $Z_{min}$ 11. The cluster has three distinct subclusters, which correspond to node numbers 1–5 in subcluster 1, 6–17 in subcluster 2 and 19–21 in subcluster 3. **B**: Graph layout of Cluster 14 showing the nodes (with PDB codes) of the cluster with node numbers corresponding to those given in (A). The protein structures of some of the nodes are shown in the figure and these are indicated by arrow from the node to the protein structure. The domains are colored in different shades and correspond to the different folds present in these proteins. For the nodes having PH domain, the figure also gives the information about the family they belong to. The segregation of nodes according to the family is clearly demonstrated by this figure.

structures although their function and the evolutionary history might be different.

### *Structure–Function Correlation in PCUG*

In addition to the structural correlation seen in PCUG, we observe that the function within a cluster is broadly conserved in many cases even if two or more than two domains are present in the cluster. It seems to suggest that domains combinations may not occur randomly. The clusters with multidomain chains bringing together two different folds also show some conservation of function. The most convincing example is the biggest cluster (cluster 4) in which there are 326 nodes and more than 15 different SCOP folds, but predominantly the cluster has carbohydrate metabolism related gene products. The functions present in the cluster can be systematically studied using the GO database.[22] GO database provides a controlled vocabulary for describing the functions of gene products. The GO database gives three different annotation schemes corresponding to the biochemical function, the biological process in which the protein is involved and the localization of the gene product. The advantages of using GO are that it is a hierarchical language, thus the functions of genes can be studied at many levels of specificity with nodes at a higher levels of specificity being children of the lower levels. The frequency plot of the GO annotation terms (GO biological process at level 4) present in the biggest cluster is given in Figure 9, from which it is clear that there is a predominant function present in this cluster (metabolism of carbohydrates). Another example of function being conserved despite the presence of two or more fold in a cluster is that of cluster 13, where most of the functions are hexose or ribose transferring enzymes, even though there are three different folds present as mentioned earlier (GO plot not shown).

Such conservation of function despite the fact that there are so many different folds in the cluster suggests that structural similarity between chains is not a random event and that the domain combinations are selected based on the functional requirements of the protein. However, it must be pointed out that clusters with diverse fold related to similar function as seen above, is not true in all cases. There are cases where the folds and the functions seen are not conserved. For example, cluster 14 consisting of PH domain and DBL homology domains seem to perform different functions. The plot of frequency of GO annotations seen in a cluster (see example in Fig. 9), which gives information regarding the distribution of functions within a cluster, reveals that functional fingerprints exist even in the case of Protein Chain Universe. Since the structural data of multidomain proteins is limited, it is difficult at present to understand fully the domain combination events through structure–function correlation of multidomain proteins.

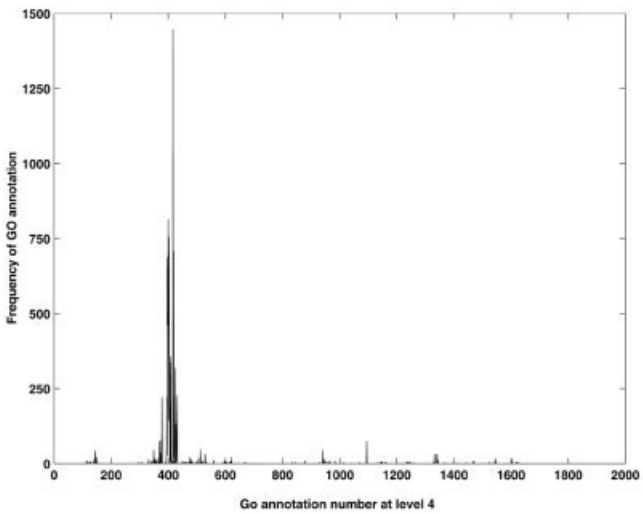


Fig. 9. Frequency plot of GO annotation (biological process) at level 4 present in biggest cluster (Cluster 4) at $Z_{min}$ 11. The annotation number is obtained from the pdb2go mapping obtained from PDB site. For each node there can be many GO terms and each GO term might be the child of many parents. Thus, the cumulative frequency of the annotation terms is higher than the number of proteins present in the cluster. The predominance of a few annotations at such a high specificity level demonstrates the fact that the cluster has proteins performing very similar tasks in the cell since they are related to very similar biological process, which is metabolism of carbohydrates.

## CONCLUSION

Protein chain universe graphs (PCUG) have been constructed using the structural similarity scores. The network properties of the graphs are analyzed as a function of the similarity scores. Further, the graph spectral algorithm is used to obtain detailed subcluster information from the completely connected graph. The nodes in a few clusters and subclusters are analyzed for structural and functional correlations. The study highlights the following points.

The network behavior of PCUG, in terms of degree distribution (number of nodes with k links) is dependent on the cutoff of the structural score used in PCUG construction. PCUG shows a scale-free behavior above a DALI Z score of 7 and the best fit with an exponent of 1.8 is obtained at a $Z_{min}$ score 11. Interestingly, the degree density (edge to node ratio normalized with respect to clique) as a function of $Z_{min}$ score shows a nice power law behavior with an exponent of 2.2.

A single numeric computation by graph spectral method yields the subcluster information in a graph. The vector components of the second lowest eigenvalue can yield information about the nature of clusters in a graph, such as the details of the number of subclusters and the density of clusters in the graph. The cluster center can be obtained from the highest vector component of the largest eigenvalue. These features can prove to be extremely useful in the analysis and annotation of large networks.

Examples are presented where the proteins are clustered together due to structural similarity or due to common domains in multidomain protein chains. In some cases, the proteins in a given cluster are involved in related functions and in some others there is no apparent functional correlation.

## SUPPLEMENTARY MATERIAL

One supplementary table has been provided, which contains all the structural and functional details of the proteins forming clusters in the PCUG at $Z_{min}$ 11.

## ACKNOWLEDGMENTS

We thank the Super computer education and research center at the Indian Institute of Science, Bangalore, India, for providing the computational facilities. KVB would like to acknowledge Council of Scientific and Industrial Research, India, for the fellowship. Support from the computational genomics initiative at IISc, funded by DBT, India, is acknowledged.

## REFERENCES

1. Wuchty S, Ravasz E, Barabási AL. The architecture of biological networks. In: Deisboeck TS, Yasha Kresh J, Kepler TB, editors. Complex systems in biomedicine. New York: Kluwer Academic Publishing; 2003.
2. Barabási AL. Linked: the new science of networks. Cambridge, MA: Persues Publishing; 2002.
3. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. Nature 2000;407: 651–654.
4. Farkas I, Jeong H, Vicsek T, Barabási AL, Oltvai ZN. The topology of transcription regulatory network in the yeast, Saccharomyces cerevisiae. Physica A 2003;318:601–612.
5. Yook SY, Oltvai ZN, Barabási AL. Functional and topological characterization of protein interaction networks. Proteomics 2004; 4:928–942.
6. Dokholyan NV, Shakhnovich B, Shakhnovich EI. Expanding protein universe and its origin from the biological Big Bang. Proc Natl Acad Sci USA 2002;9922:14132–14136.
7. Wuchty S. Scale-free behavior in protein domain networks. Mol Biol Evol 2001;18:1694–1702.
8. Kannan N, Vishveshwara S. Identification of side-chain clusters in protein structures by a graph spectral method. J Mol Biol 1999;2922:441–464.
9. Vishveshwara S, Brinda KV, Kannan N. Protein structure: insights from graph theory. J Theor Comput Chem 2002;1:187–211.
10. Sistla RK, Brinda KV, Vishveshwara S. Identification of domains and domain interface residues in multi-domain proteins from graph spectral method. Proteins 2005;59:616–626.
11. Apic G, Gough J, Teichmann SA. Domain combinations in Archeal, Eubacterial, and Eukaryotic proteomes. J Mol Biol 2001;310:311–325.
12. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
13. Holm L, Sander C. Touring protein fold space with DALI/FSSP. Nucleic Acid Res 1998;26:316–319.
14. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
15. Altschul SF, Madden TL, Schaffer 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.
16. West DB. Introduction to graph theory. Prentice-Hall of India Private Limited; 2000.
17. Gansner ER, North SC. An open graph visualization system and its applications to software engineering. Softw Pract Exper 1999; 00S1:1–5.
18. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res 2002;30:264–267.
19. Deeds EJ, Dokholyan NV, Shakhnovich EI. Protein evolution within a structural space. Biophys J. 2003;855:2962–2972.
20. Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. Nature 2002;420:218–223.
21. Dietmann S, Park J, Notredame C, Heger A, Lappe M, Holm L. A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. Nucleic Acids Res. 2001;29:55–57.
22. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genet 2000;25:25–29.