International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 6 (2014) pp. 305 - 314

Graph Centrality Analysis of Structural Ankyrin Repeats

Broto Chakrabarty and Nita Parekh

Center for Computational Natural Sciences and Bioinformatics,
International Institute of Information Technology, Hyderabad, India
{broto.chakrabarty@research., nita@}iiit.ac.in

***Abstract*: In recent studies it has been shown that graph representation of protein structures is capable of capturing the 3-dimensional fold of the protein very well, thus providing a computationally efficient approach for protein structure analysis. Centrality measures are generally used to identify the relative importance of a node in the network. Here we demonstrate a novel application of centrality analysis: to identify tandemly repeated structural motifs in 3-d protein structures. This is done by analyzing the profile of various centrality measures in the repeat region. The comparative analysis of five centrality measures based on local connectivity, shortest paths, principal eigen spectra and feedback centrality is presented on proteins containing contiguous ankyrin structural motifs to identify which centrality measure best captures the repetitive pattern of ankyrin. We observe that principal eigen spectra of the adjacency matrix and Katz status index, both exhibit a distinct profile for the ankyrin motif capturing its characteristic anti-parallel helix-turn-helix fold. No such conserved pattern was observed in the repeat regions of equivalent random networks, suggesting that the conserved pattern arises from the 3d fold of the structural motif.**

***Keywords*:** Ankyrin repeat, graph theory, protein contact network, centrality measures.

I. Introduction

A protein fold is governed by covalent and non-covalent interactions between its residues. These interactions are captured in protein contact network (PCN) by computing the Euclidean distance between the amino acids and drawing an edge between the residues lying within a pre-defined threshold distance (~ 7Å). The connection topology of this network reflects the 3-dimensional fold of the protein molecule and provides an alternative computational approach for structural analysis of proteins. Several techniques for such structural analysis exist, such as the analysis of the global network structure, network motifs, clustering and network centralities. Network centralities are used to rank elements (residues) and identify key elements in a network. The idea of centrality was first introduced by Bavelas (1948) in an attempt to understand communication in small groups. Since then, the study of centrality has been used to address different problems such as political integration, design organizations, communication paths, social influence, etc. [1]. Since nodes having similar neighborhood are expected to have similar centrality values, subgroups of amino acids in a protein with similar 3-d fold are expected to exhibit similar pattern in their centrality profile. With this aim here we carry out an analysis of various centrality measures to identify contiguous structural repeats.

Repetition of a super secondary structure within a protein is a common phenomenon observed in about 14% of proteins [2]. The copy number of these repeats and their arrangement account for large number of structural and functional roles such as protein transport, protein-complex assembly, and protein regulation. Different repeats such as leucine-rich repeat (LRR), ankyrin repeat (ANK), tetratricopeptide repeat (TPR), etc. have been defined based on the repetition and arrangement of the specific super secondary structure. Here, we present our analysis on proteins containing ankyrin repeat, which is a helix-turn-helix motif about 30-34 amino acids long, and exclusively functions to mediate protein-protein interactions such as transcription initiation, cell-cycle regulation, cytoskeletal regulation, ion transport and signal transduction. It is one of the most frequently observed protein motifs in nature and their abundance makes it desirable to identify them to understand their biological functions.

The problem of identification of repeats in protein has been addressed by various sequence alignment [3] [4] and profile based methods [5]. The sequence-based methods are generally reliable when the sequence conservation is high within individual repeat copies. However, it has been observed that the sequence conservation between individual repeating units can be very low (~15%). Thus, with the increase in the number of available protein structures, it is desirable to design structure based methods to identify repeats in proteins. Methods such as OPAAS [6] and DAVROS [7] implement self-structural alignment of proteins, while Swelfe [8] and ProSTRIP [9] implement dynamic programming on sequence of α angles derived from dihedral angle for the identification of repeats at the structure level. IRIS implements structural alignment with its database of internal repeat units if no confident results are obtained from the sequence based methods [10]. The structure-based methods, being computationally very intensive, here we investigate graph-based approach for the identification of structural repeats in proteins. Centrality analysis has been shown to be a valuable method for the structural analysis of biological networks. Here we discuss and compare the profile of various centrality measures in the repeat regions to assess which centrality best reflects the repetitive pattern of the structural

motif.

II. Protein Contact Network

A Protein Contact Network (PCN) is a representation of the protein structure in the form of a mathematical entity, a graph. The idea is to capture the interactions responsible for maintaining secondary structures and stabilizing the 3-dimensional fold of the protein [11]. It is well-known that a protein structure is governed to a large extent by non-covalent interactions. The non-bonded interactions such as van der Waals forces and hydrogen bonds are responsible for the unique three-dimensional fold of the proteins. These interactions are constraint by the spatial proximity of other atoms, and in the graph representation this is realized by considering a threshold distance between the amino acid residues in the three-dimensional space. The coordinates of amino acids in a protein structure are extracted from PDB [12] record file to compute the distances between the atoms and draw links between them based on their spatial proximity to capture the non-bonded interactions.

In this study, we construct PCN as an undirected graph $G = (V, E)$ which consists of a finite set V of vertices ($n = |V|$) where C_α atom of each amino acid is considered as a vertex. Two vertices u and v are connected by an edge $e = (u, v) \in E$ if the Euclidean distance between the C_α atoms of the amino acids represented by the vertices u and v is within 7Å (R_c). In Fig. 1 (a) and (b) are shown the 3-dimensional structure of a designed ankyrin repeat protein, 1N0R and the corresponding protein contact network respectively. It may be noted that the interactions between two anti-parallel helices of the ANK motif are very clearly captured in the protein contact network.

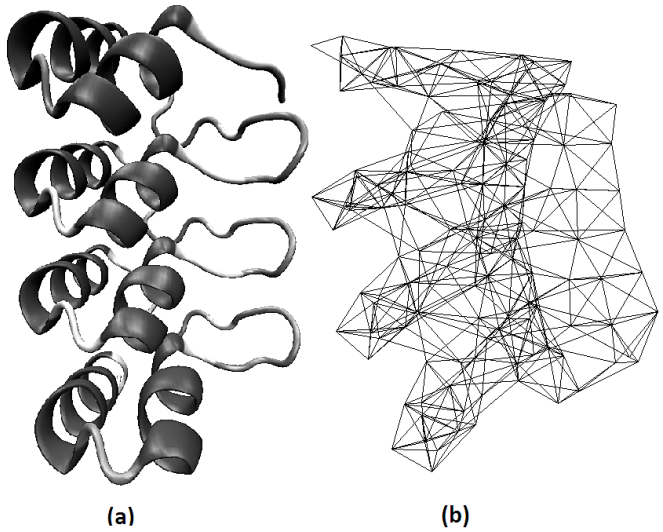

Figure 1. (a) The 3-d structure, and (b) protein contact network of protein 1N0R.

The connectivity information in a graph is mathematically represented by a $n * n$ adjacency matrix, whose elements, A_{uv}, take a value '1' if the nodes u and v are directly connected to each other (i.e., if $d_{uv} \leq R_c$ and $u \neq v$), '0' otherwise. Here, $d_{uv} = \sqrt{\{(x_u - x_v)^2 + (y_u - y_v)^2 + (z_u - z_v)^2\}}$ is the Euclidean distance between every (u,v) pair and (x,y,z) are the coordinates of the C_α atoms extracted from PDB record using a python script.

III. Dataset

For the analysis, a non-redundant set of proteins containing ankyrin repeats is constructed. The dataset contains both designed and natural proteins. The designed ankyrin proteins are obtained from the SCOP database [13] and natural proteins containing ankyrin repeats are obtained by keyword search from Pfam [14] and PROSITE [15] databases. The structures of these proteins are obtained from Protein Data Bank (PDB). To remove redundancy, only high resolution structures (< 3Å) corresponding to a unique UniProt entry are considered for the analysis.

IV. Method

In PCN each amino acid contributes to the connectivity of the network. The property of a node in a network is analyzed by quantitative measure called centrality which is a mathematical function defined to rank the vertices in the network. Centrality can be used to address many problems in complex networks. For example, in a social network, the influential individuals can be identified as those who are connected to a large number of individuals or those connected to individuals with large number of connections.

Large number of graph centrality measures has been defined for the analysis of various topological networks, including biological networks. Each centrality measure captures a specific property of the graph. The structural repeats in a protein contain repetition of a super secondary structure several times within the protein. It is expected that the structural repeats may exhibit similar topological properties which can then be exploited for their identification. With this aim here we present an analysis of various centrality measures to identify which of these best capture the repetitive pattern of structural motifs in proteins.

The graph centrality measures used in this analysis are briefly discussed below for a small undirected representative graph in Fig. 2. The graph has 10 nodes represented by numbered circles and 11 edges between these nodes represented by straight lines connecting the nodes. The subgraph formed by nodes 1, 2, 3 and 4 has similar topology to the subgraph formed by nodes 6, 7, 8 and 9, except that one extra node is connected to node 9.

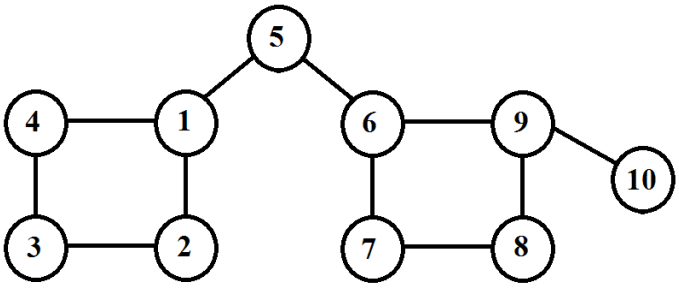

Figure 2. A schematic representation of an undirected graph with 10 nodes and 11 edges.

The connectivity of a graph is mathematically represented by adjacency matrix, A, given below for the graph in Fig. 2:

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \quad (2)$$

where, A_{uv} is '1' if nodes *u* and *v* are connected by an edge, '0' otherwise. The graph centrality measures studied here can be classified as: local centrality, distance based centrality, eigen spectra centrality and feedback centrality. A local centrality is computed based on the immediate neighbors of the node and gives a local measure of the importance of the node. A distance based centrality considers the shortest path distances between various nodes in the network to define its importance. The eigen spectra centrality is derived from various matrices used to represent the graph such as adjacency and Laplacian matrices. Feedback centrality of a node is defined recursively from the centrality of its adjacent nodes. These centralities are defined below for the representative network in Fig. 2, and the magnitude of the measures for each node is summarized in Table 1.

A. Local centrality measures

A local centrality measure is defined by the immediate neighbors of a node and depicts the importance of the node in immediate environment.

1) Degree

The most simple centrality measure is the degree centrality, C_d (*u*), defined as the number of nodes to which the node *u* is directly connected. It is defined as:

$$C_d(u) = \sum_{v \in V} A_{uv} \quad (2)$$

where, A_{uv} is '1' if the residue u is in spatial proximity to the residue *v* in a protein contact network, '0' otherwise. For instance, the residues within the hydrophobic core of the protein are likely to have a high degree compared to the residues present at the surface of the protein or in the loop region which have relatively less intra-molecular interactions. For the graph in Fig. 2, nodes 1, 6 and 9 have high degree (= 3) (Table 1), and their removal will lead to disjoint clusters.

2) Clustering coefficient

The clustering coefficient of a node u is a measure of connectivity of its neighbors and is given by:

$$C_{cc}(u) = \frac{\frac{1}{2}\sum_{v \in V}\sum_{w \in V} A_{uv}A_{uw}A_{vw}}{{}^{d}C_2} \quad (3)$$

where, *A* is the adjacency matrix, *v* and *w* are neighbors of *u*, *d* is the number of nodes connected to the node *u* and *V* is the vertex set. For the graph in Fig. 2, the clustering coefficient of all the nodes is zero since none of the neighbors of a node are directly connected to each other. However, if we introduce an edge between nodes 4 and 2, then the clustering coefficient of node 3 will be 1 as it has two neighbors (node 4 and 2) which will be connected to each other, while nodes 2 and 4 each have 3 neighbors, of which two pairs would be connected, and the clustering coefficient for these nodes would be 2/3 = 0.67.

Node	C_{degree}	C_{btw}	C_{cl}	A_{levc}	L_{ssevc}	C_K
1	3	18.5	0.048	0.345	-0.404	2.211
2	2	3.5	0.037	0.232	-0.48	1.854
3	2	0.5	0.03	0.198	-0.506	1.792
4	2	3.5	0.037	0.232	-0.48	1.854
5	2	20	0.053	0.344	-0.21	1.961
6	3	21	0.053	0.461	0.006	2.292
7	2	3	0.04	0.331	0.066	1.889
8	2	1	0.034	0.314	0.119	1.874
9	3	11	0.043	0.405	0.161	2.203
10	1	0	0.032	0.173	0.179	1.47

Table 1. Centrality measures of the undirected network.

B. Distance based centrality measures

The distance based centralities are defined by the shortest path distances between the nodes. These measures depict the transfer of information in a network.

1) Betweenness

The idea of betweenness as a centrality is based on the observation that an important node will lie on a large number of paths between other nodes in the network, i.e. nodes that can control the information/communication flow through the network [16]. The betweenness centrality of a node u is the number of geodesics going through it and is defined as:

$$C_{btw}(u) = \sum_{s \neq u \in V} \sum_{t \neq u \in V} \sigma_{st}(u)/\sigma_{st} \quad (4)$$

where, σ_{st} is the number of shortest paths from residue *s* to *t*, and $\sigma_{st}(u)$ is the number of shortest paths from *s* to *t* that pass through *u*. Betweenness centrality helps in identifying nodes that make the most contribution in transmission flow in the network. In the analysis of protein contact networks, the betweenness centrality is shown to be useful in identifying and characterizing residues that are important for folding [17]. In an earlier work, we have shown the usefulness of this measure in the identification of ARM/HEAT structural repeats in proteins [18].

For the graph in Fig. 2, the nodes 1, 5 and 6 are centrally located with maximum number of shortest paths passing through these nodes. Thus, these nodes show significantly high values for betweenness centrality. Only one shortest path passes through node 3, i.e. between nodes 2 and 4, which has an alternate shortest path through node 1. Thus, out of two possible paths between nodes 2 and 4, one passes through 3 which is shown by the betweenness value of 0.5. Since, no shortest path passes through node 10, its betweenness value is 0. High degree nodes in general have high betweenness values as many shortest paths may pass through them. However, a high betweenness node need not always be a high degree node, for example node 5, because of its topological location in the graph.

2) Closeness

The closeness centrality of a residue *u* measures how easily all other residues in the graph can be reached from it, and is defined as:

$$C_{cl}(u) = (n-1)/\sum_{v \epsilon V} d(u,v) \quad (5)$$

where *n* is the size of the network and *d(u,v)* is the shortest path distance between the pair of residues *u* and *v*. The mathematical formula was derived by Beauchamp [1] in 1965 defining a node as important if it is close to all other nodes and can transfer information quickly. It is observed in a protein contact network that the closeness centrality is typically high for active site, ligand-binding and evolutionary conserved residues [19]. For the representative graph in Fig. 2, nodes 5 and 6 have highest equal closeness centrality indicating that these nodes have minimum cumulative shortest path distance from all other nodes and are very well connected to all other nodes. The total shortest path distance of both the nodes 5 and 6 to all other nodes is 19, which makes their magnitude equal.

C. Eigen spectra centrality

The centrality of a node may also depend on the centrality of the nodes it is connected to. This information is captured by the eigen spectra of the connectivity matrix of a graph, such as adjacency matrix and Laplacian matrix.

1) Eigen Spectra of Adjacency Matrix

The eigenvector components corresponding to the principal eigenvalue of the adjacency matrix have been shown to provide information on the structure and topology of the graph [20] [21]. It not only captures the connectivity of a node but also that of nodes adjacent to it, and nodes adjacent to its neighbors, and so on. Thus, the graph spectral analysis is useful in identifying the connectivity pattern of a group of nodes, clusters, in the network.

If for the i^{th} node, the centrality score is proportional to the sum of the scores of all nodes which are connected to it, then

$$x_i = \frac{1}{\lambda}\sum_{j \in M(i)} x_j = \frac{1}{\lambda}\sum_{j=1}^{N} a_{ij}x_j \quad (6)$$

where $M(i)$ is the set of nodes that are connected to the i^{th} node, N is the total number of nodes, and λ is a constant. In vector notation this is written as

$$x = \frac{1}{\lambda} Ax, or\ Ax = \lambda x \quad (7)$$

For the representative graph in Fig. 2, the eigenvector centrality, A_{levc}, of nodes 2 and 4 are equal in magnitude as these nodes share same neighbors (Table 1). Thus, the contribution of the neighbors, and neighbor's neighbors and so on, are same for both these nodes. Similarly, we expect the corresponding residues in contiguous structural repeats to exhibit similar A_{levc} values as they have similar 3d topology. In a preliminary study, we did observe a conserved profile of A_{levc} in the repeat regions [22].

2) Eigen Spectra of Laplacian Matrix

Another extensively studied matrix of a graph is the Laplacian matrix, L, which is a modified form of the adjacency: $L = D - A$, where D is a diagonal matrix containing the degree information of the nodes. For the graph in Fig. 2, the Laplacian matrix is

$$L = (-1)\begin{bmatrix} -3 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -3 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & -3 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \quad (8)$$

It has been shown that the eigenvector components corresponding to the second smallest eigenvalue (L_{ssevc}) captures the clustering information; i.e. nodes that belong to a particular cluster have the same sign and nodes with similar neighborhood connectivity pattern have same magnitude of L_{ssevc}. For the representative graph in Fig. 2, two distinct clusters are obtained. The nodes 1, 2, 3 and 4 have similar magnitude of L_{ssevc} as represented in Table 1 which forms one cluster, and the other cluster is formed by the nodes 6, 7, 8, 9 and 10 with another set of similar L_{ssevc} values. Here we expect the L_{ssevc} values of the residues in a repeat unit to be similar.

D. Feedback centrality

Feedback centrality of a node is based on the assumption that a node is more central, if its neighbors have high centrality values.

1) Katz Status Index

It computes the relative influence of a vertex u within a network by measuring not only the number of the immediate neighbors but also all other vertices in the network that connects to the vertex u through its immediate neighbors and is given by

$$C_k(u) = \sum_{k=1}^{\infty}\sum_{v=1}^{n} \alpha^k (A^k)_{vu} \quad (9)$$

It incorporates indirect influence of other nodes through an attenuation factor, $\alpha > 0$, that reduces the contributions from nodes at increasing lengths on the node under consideration [23]. Unlike other centrality measures that consider only the shortest path between a pair of nodes, Katz centrality measure takes into account the total number of paths between a pair of vertices. For the representative graph, node 6 is having the highest C_k value as it has a highest degree and its neighbors are well connected to other nodes in the network and have high C_k values.

V. Analysis of Centrality Measures

Here we provide an analysis of different categories of centrality measures to see which of the measures based on

local connectivity information, shortest-path distances, feedback or spectral analysis of the connectivity matrix is able to better capture the repetitive 3-dimensional structural topology of ankyrin repeats. We first present our analysis of the protein contact network (PCN) of designed ankyrin repeat protein, 1N0R (Fig. 1), which contains four consecutive repeats. Using sequence-based repeat identification tool, RADAR [3], the boundaries of the four consecutive repeats is identified as shown in Fig. 3. The centrality measures degree, clustering coefficient, betweenness, principal eigen spectra of adjacency matrix and Katz Status index are plotted in Fig. 4 (a), (c), (e), (g) and (i) respectively. The profile of these five centrality measures for the individual repeat regions are superimposed in Fig. 4 (b), (d), (f), (h) and (j) respectively. The vertical dotted and solid lines correspond to the start and end of the repeat boundaries, predicted by RADAR output.

No. of Repeats	Total Score	Length	Diagonal	BW-From	BW-To	Level
4	244.11	31	31	11	41	1

1-	33 (66.52/43.10)	NGRTPLHLaaRNGHLEVVKLLLEAGADVNAKDK
34-	66 (66.52/43.10)	NGRTPLHLaaRNGHLEVVKLLLEAGADVNAKDK
67-	99 (66.52/43.10)	NGRTPLHLaaRNGHLEVVKLLLEAGADVNAKDK
100-	125 (44.55/26.52)	NGRTPLHLaaRNGHLEVVKLLLEAGA.....

Figure 3. RADAR output of the protein 1N0R.

From Fig. 4 (a) and 4 (c) it is clear that both the degree and clustering coefficient profiles for the individual repeat regions is very similar, with the pattern being well-conserved in the core repeat region. This is further confirmed by overlapping these profiles for the individual repeat copies in Fig. 4 (b) and 4 (d) respectively. The profile of the degree in the repeat regions is better conserved compared to clustering coefficient. The profile of the distance based centrality measure, betweenness, is shown in Fig. 4 (e)-(f) for protein 1N0R. A repetitive pattern in the profile in this centrality measures is observed to be well-conserved in the core of the repeat region, as is clear in Fig. 4 (f) obtained on overlapping the profiles for the individual repeat regions. The first helix of the repeat units in ankyrin is present at the core of the protein and the second helix is at the surface away from the core. Consequently, most of the shortest paths in the network pass through the residues of the first helix as compared to the residues of second helix. This property is quantitatively reflected by the high values of betweenness for the first helix residues from 5 to 12 and low values for the second helix residues from 14 to 24 in Fig. 4 (e). However, compared to degree which has a time complexity of $O(|V|)$, the complexity of computing betweenness centrality is $O(|V||E|)$, where $|V|$ and $|E|$ are the total number of vertices and edges in the graph, making it computationally expensive [24]. The eigen spectra of the matrices associated with a graph, i.e., the adjacency and Laplacian matrices are known to capture very well the topology of the graph and identify clustering patterns [20]. Here we investigate their efficacy in identifying tandemly repeated structural motifs. In Fig. 4 (g) is plotted the principal eigenvector components of the adjacency matrix, A_{levc} and the plot showing the overlap of A_{levc} profiles for the individual repeat regions in Fig. 4 (h). A very clear pattern with two peaks corresponding to the two helices for each individual repeats is observed. Since the vector components contain contribution from its neighbors, neighbor's neighbors, and so on, the centrally located repeats show very prominent pattern tapering on either sides of the core of the repeat region, though retaining the overall shape of the profile at both ends. The time complexity of computing the eigenvector centrality is ~ $O(V^2)$ for a sparse matrix and $O(V^3)$ if the network is dense, where V is the number of vertices.

It is known that the eigenvector components corresponding to the second smallest eigenvalue of the Laplacian matrix, L_{ssevc}, captures the clustering information. The L_{ssevc} values of the residues belonging to each repeat motif were analyzed and found to be within one standard deviation from the mean.

The plots for the feedback centrality, Katz status index is given in Fig. 4 (i)-(j). The Katz status centrality exhibits pattern very similar to the principal eigen spectra of the adjacency matrix in Fig. 4 (g). This is not surprising since the Katz status index takes into consideration the influence of all indirect links through intermediates, similar to eigen vector component. It is supposed to be most suitable in the case of directed acyclic graphs where eigen spectra analysis fails [25]. The time complexity of Katz index is limited by matrix inversion step which is $O(V^3)$ for V vertices, with faster versions of the algorithm being of $O(V+E)$, E being the number of edges. From the above analysis we observe that the three centrality measures, viz., degree, eigenvector centrality and Katz status index capture very well the repetitive pattern. However, the A_{levc} profile is more prominent and well-conserved than the degree profile in the repeat regions. This is not surprising since eigenvector centrality is like a recursive version of degree centrality; it is large for a node if either it has many neighbors and/or it has important neighbors. That is, it captures not only the connectivity of a node, but its neighbor's connectivity, neighbor's neighbor's connectivity, and so on. The Katz status index is defined as a generalization of degree centrality and can be written as a variant of eigenvector centrality and hence its profile in the repeat regions is very similar to that of A_{levc}. Its profile in the repeat regions is even better conserved than both the degree and A_{levc} profiles, as it captures not only the direct links to a node, but all indirect links to it through an attenuation factor, $\alpha > 0$, that reduces the contributions from nodes at increasing lengths from it. Also, it takes into account the total number of paths between a pair of vertices instead of only shortest path between a pair of nodes, unlike other centrality measures.

These observations suggest that one may use degree, principal eigen vector components of the adjacency matrix, A_{levc} or Katz status index for the identification of structural repeats. Since the variation in the loop region for A_{levc} is much lower compared to the other two, identifying repeat boundaries is most reliable with the eigen spectra of adjacency matrix.

The protein 1N0R being artificially designed based on the consensus ANK motif, it is not surprising that majority of the centrality measures analyzed are able to capture the repetitive topology of the ANK motif. We next investigate the profiles of these centrality measures for a natural protein, 3EHQ, comprised of repeat regions and non-repeat regions. Fig. 5 (a) and (b) show the 3-d structure and corresponding PCN of the protein, 3EHQ which has 3 consecutive ANK motifs in the

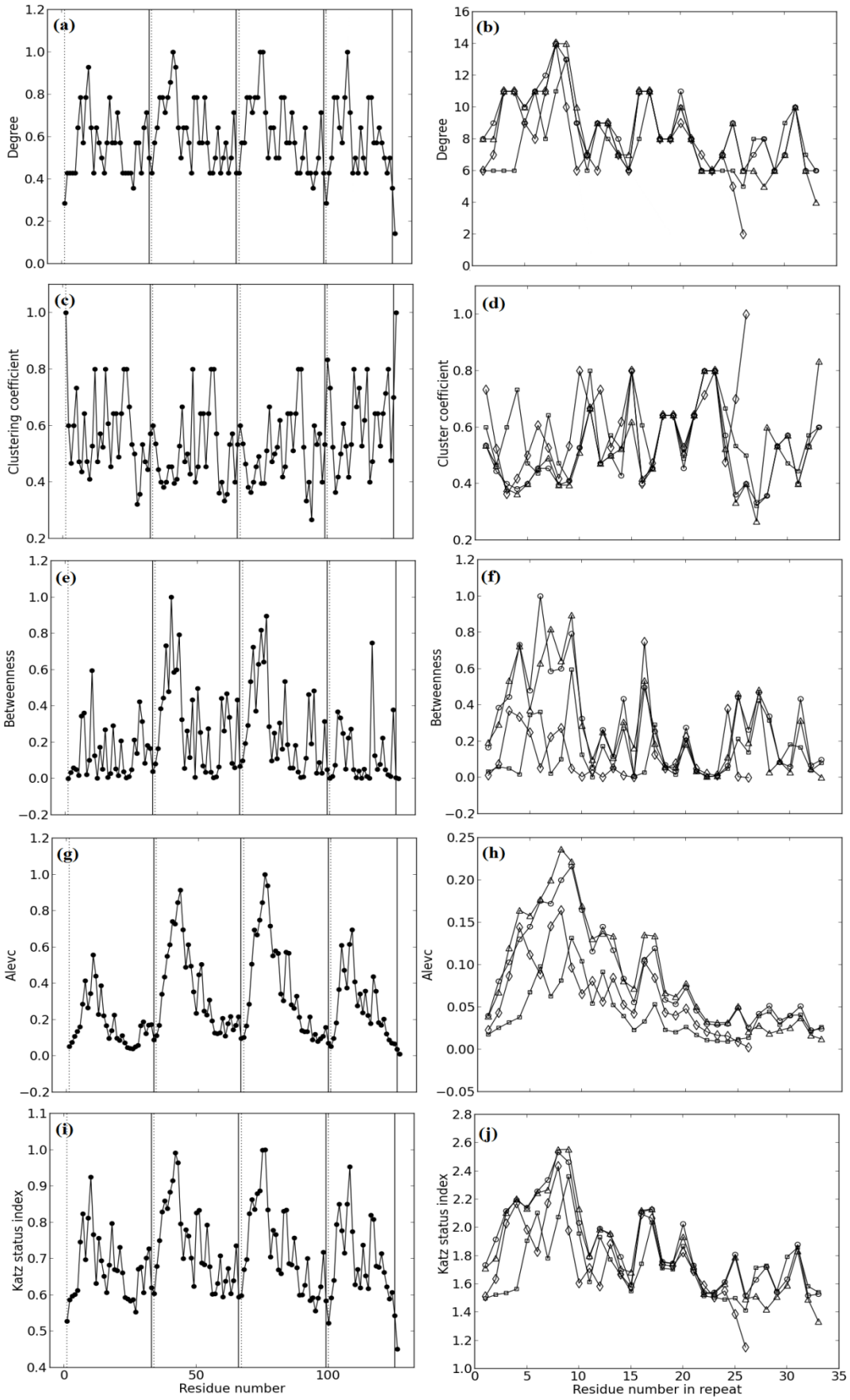

Figure 4. The centrality measures computed for each residue of the protein 1N0R shown: (a) Degree, (c) Clustering coefficient, (e) Betweenness, (g) principal eigenvector of adjacency matrix (A_{levc}) and (i) Katz status index. Start and end of each repeat is taken from RADAR output and are indicated by dotted and solid lines respectively. In (b), (d), (f), and (h), the respective centrality measures are plotted by overlapping centrality profiles of repeat regions.

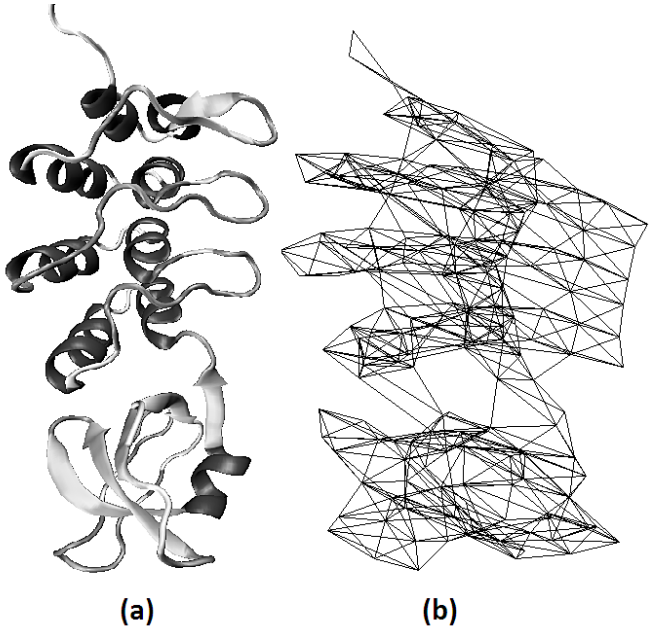

Figure 5. (a) The 3-d structure, and (b) protein contact network of protein 3EHQ.

C-terminal from 72 to 168. The plots of five centrality measures along with the overlapping repeat regions for 3EHQ are given in Fig. 6.

The degree centrality and clustering coefficient in Fig. 6 (a)-(b) and (c)-(d) respectively, shows a conserved pattern in the repeat regions but the start boundary is not identifiable. The distance based measure betweenness in Fig. 6 (e)-(f) show a conserved pattern for two repeat regions similar to the designed protein. The C-terminal repeat does not show a clear pattern for the distance based centralities. The eigen spectra based centralities A_{levc} is shown in Fig. 6 (g)-(h). The profile of A_{levc} shows a clear pattern with two peaks corresponding two helices in each of the three repeat units. A significant difference in the magnitude of A_{levc} between repeat region and non-repeat region is observed in Fig. 6 (g), which is not so in case of the feedback centrality Katz status index, although it shows a conserved pattern in the repeat regions (Fig. 6 (i)). The variation in the loop regions of natural protein, 3EHQ, for A_{levc} is much lower compared to the feedback centralities, as observed in the case of designed protein, 1N0R and helps in the identification of the repeat boundaries.

A clear 2-peak pattern in the principal eigen spectra is observed in the repeat regions of the proteins 1YCS, 1AWC, 1NFI and 1N11, shown in Fig. 7. The vertical dotted and solid lines in this case correspond to the start and end boundaries of each repeating unit based on the annotation provided in the UniProt database. Pattern of two peaks for the two anti-parallel helices is clearly observed for individual ANK motifs in each of these proteins, clearly suggesting that the eigen spectra analysis is reliable for identifying tandem ANK repeats.

VI. Analysis of Random Network

The pattern observed in the principle eigen spectra of the adjacency matrix for the structural repeat regions is observed due to similar connectivity pattern of the residues in these regions. Next we randomize the protein contact network to see if the repetitive pattern in eigenvector centrality is lost on randomizing the connection topology. The randomization of the network is done by keeping the number of nodes and edges constant as in the original network and also retaining the backbone connectivity of the protein chain. Thus every non-backbone edge is randomly assigned to two randomly chosen nodes. That is, the backbone chain of the protein structure is kept constant but the 3-d conformation is altered. For the designed protein 1N0R and the natural protein 3EHQ, 100 randomizations were performed and the principal eigen spectra in the repeat region analyzed. The A_{levc} plot for one such random configuration for proteins 1N0R and 3EHQ is shown in Fig. 8 (a) and (c) respectively and the plots obtained on overlapping the A_{levc} profiles for the repeat regions are given in Fig. 8 (b) and (d) respectively. The pattern conserved in the repeat regions for 1N0R (Fig. 4 (h)) is lost in this case as seen Fig. 8 (b). Similarly, the conserved pattern observed in the repeat regions of 3EHQ in Fig. 6 (h) is not observed in the randomized counterpart in Fig. 8 (d). This confirms that the A_{levc} pattern observed in the repeat regions of ANK proteins is specific to this repeat type.

VII. Conclusion

The representation of protein structures as networks provide insight into the complex 3-dimensional topological features of proteins. By representing protein as a graph, it is reduced to a mathematical entity on which computationally efficient algorithms can be used to identify and analyze important structural features. In this study we consider an important pattern recognition problem, *viz.*, identifying tandemly repeated structural motifs using graph centrality measures. Different centrality measures capture the importance of a node based on a different concept. A comparative analysis of five centrality measures has been presented here to analyze and identify for the most widely observed structural motif in proteins, the Ankyrin motif. The spectral analysis of the adjacency matrix and the Katz status index are observed to be most reliable of all the measures analyzed here. The advantage of the proposed graph based approach is that no domain information is used for the identification of tandem structural repeats. Compared to traditional approaches, the graph based approaches are also computationally very efficient since no sequence/structure-based alignment is required. The limitation of the proposed method is that it is qualitative for defining accurately the boundaries of the repeats domain information such as architecture of secondary structural elements is desirable.

Acknowledgment

The authors acknowledge the support of CSIR funding for this work (37(1468)/11/EMR-II).

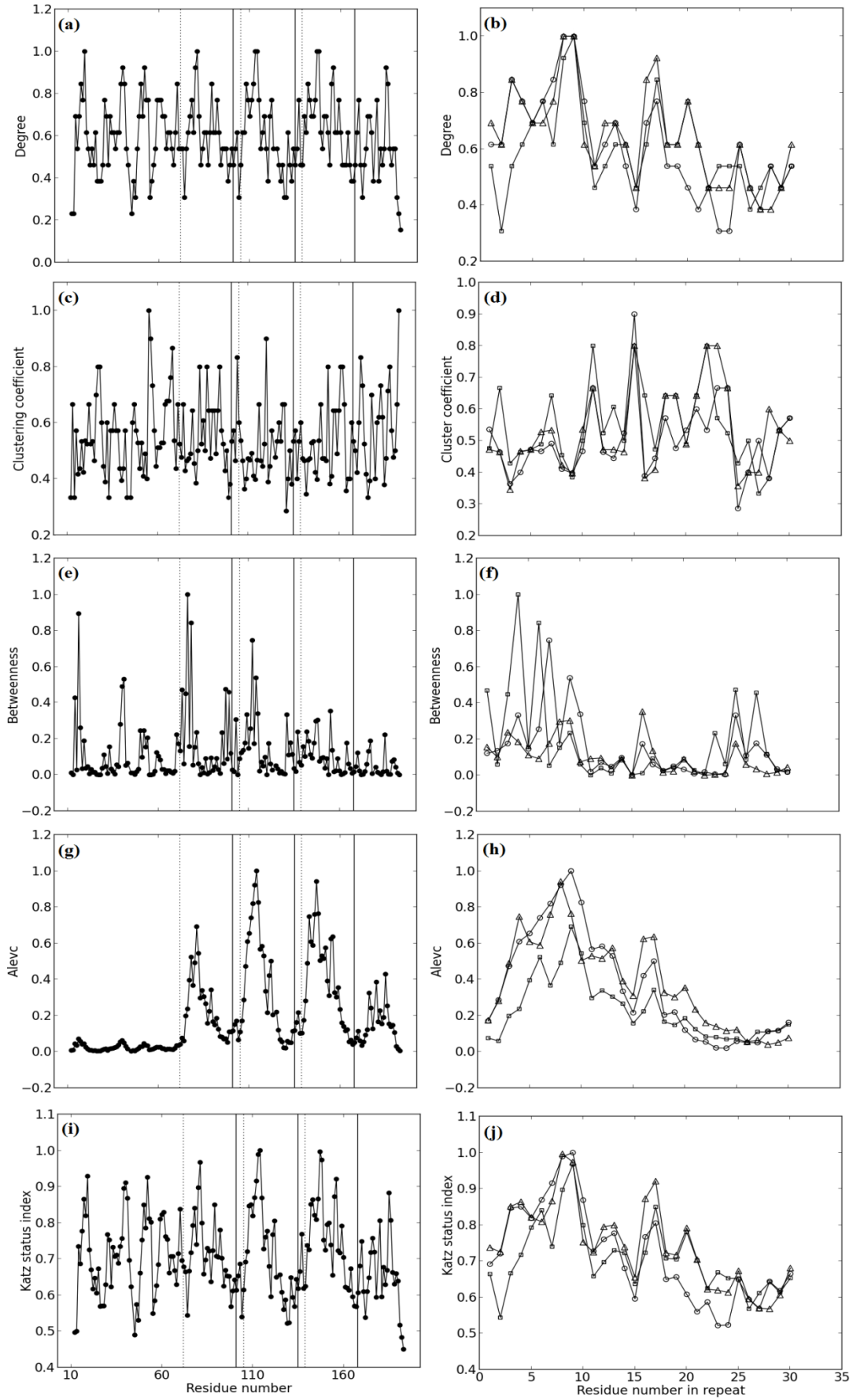

Figure 6. The centrality measures computed for each residue of the protein 3EHQ shown: (a) Degree, (c) Clustering coefficient, (e) Betweenness, (g) principal eigenvector of adjacency matrix (A_{levc}) and (i) Katz status index. The start and end of each repeat is taken from UniProt annotation and are indicated by dotted and solid lines respectively. In (b), (d), (f), (h) and (j), the respective centrality measures are plotted by overlapping the repeat regions.

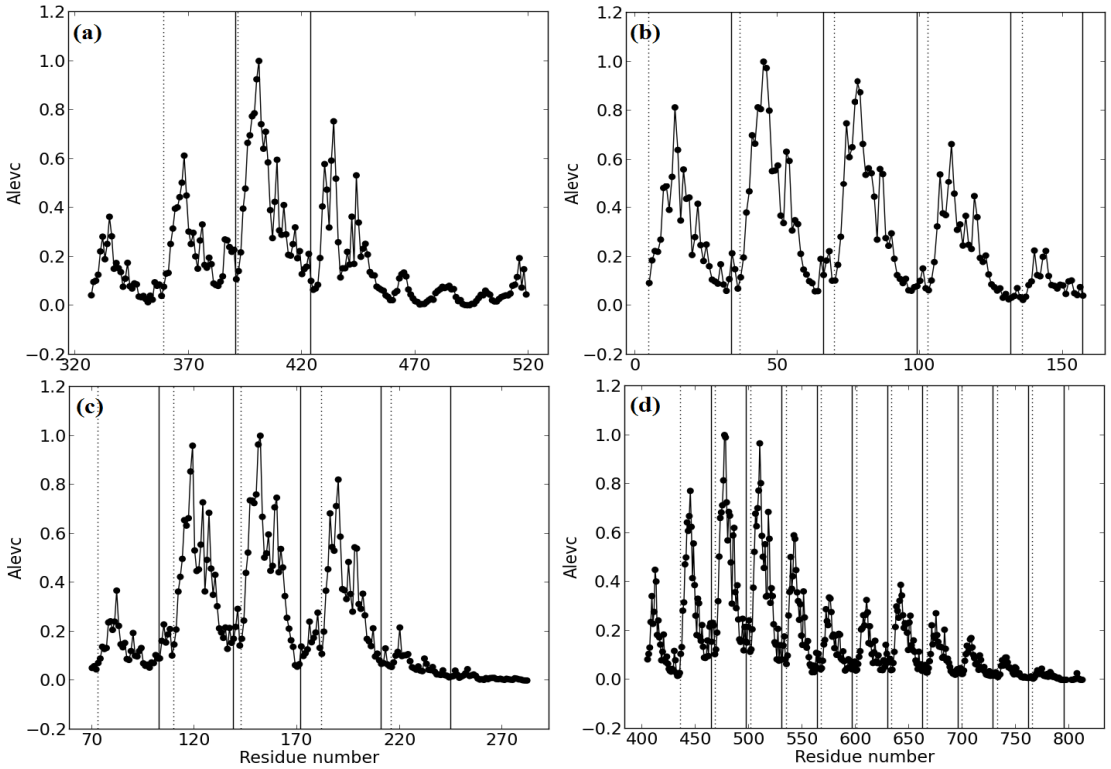

Figure 7. Principal eigen spectra of adjacency matrix, Alevc for proteins: (a)1YCS, (b)1AWC, (c)1NFI, and (d)1N11 shown. A clear two-peak pattern observed in each case in the repeat regions.

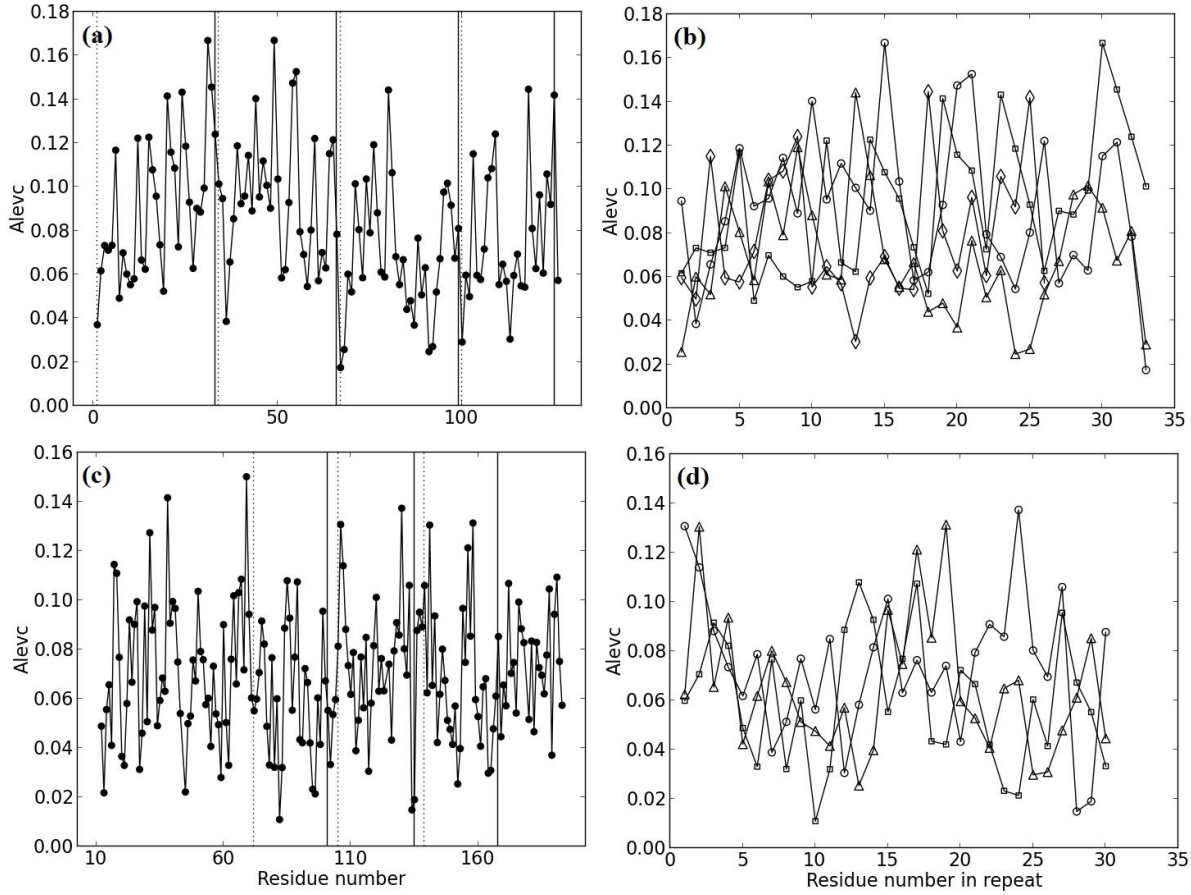

Figure 8. The plot of principal eigenvector of adjacency matrix (Alevc) for randomized networks of (a) 1N0R, and (c) 3EHQ shown. The start and end of the repeat regions are indicated by dotted and solid lines respectively. In (b) and (d), the respective centrality measures are plotted by overlapping the repeat regions.

References

[1] L. C. Freeman. "Centrality in Social Networks Conceptual Clarification," *Social Networks*, 1, pp. 215-239, 1978.

[2] M. A. Andrade, C. Perez-Iratxeta, and C. P. Ponting. "Protein Repeats: Structures, Functions, and Evolution," *J. Struc. Biol.*, 134, pp. 117-131, 2001.

[3] A. Heger, and L. Holm. "Rapid Automatic Detection and Alignment of Repeats in Protein Sequences," *Proteins: Structure, Function and Genetics*, 41, pp. 224-237, 2000.

[4] R. Szklarczyk and J. Heringa. "Tracking repeats using significance and transitivity," *Bioinformatics*, 20, pp. i311–i317, 2004.

[5] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, 25, 17, pp. 3389-3402, 1997.

[6] E. S. Shih, R. R. Gan, and M. Hwang. "OPAAS: a web server for optimal, permuted, and other alternative alignments of protein structures," *Nucleic Acids Research*, 34, pp. W95-W98, 2006.

[7] K. Murray, W. Taylor, and J. Thornton. "Toward the Detection and Validation of Repeats in Protein Structure," *PROTEINS: Structure, Function, and Bioinformatics*, 57, 2, pp. 365-380 2004.

[8] A. Abraham, E. Rocha, and J. Pothier. "Swelfe: a detector of internal repeats in sequences and structures," *Bioinformatics*, 24, 13, pp. 1536-1537, 2008.

[9] R. Sabarinathan, R. Basu, and K. Sekar. "ProSTRIP: A method to find similar structural repeats in three-dimensional protein structures," *Computational Biology and Chemistry*, 34, pp. 126-130, 2010.

[10] H. Kao. "A Comprehensive System for Identifying Internal Repeat Substructures of Protein," In *Proceedings of International Conference on Complex, Intelligent and Software Intensive Systems, IEEE*, pp. 689-693, 2010.

[11] R. Saidi, M. Maddouri, and E. M. Nguifo. "Comparing graph-based representations of protein for mining purposes," In *Proceedings of KDD-09 Workshop on Statistical and Relational Learning in Bioinformatics*, pp. 35-38, 2009.

[12] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne. "The Protein Data Bank," *Nucleic Acids Research*, 28, 1, pp. 235-242, 2000.

[13] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. "SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures," *J. Mol. Biol*, pp. 536-540, 1995.

[14] E. L. Sonnhammer, S. R. Eddy, and R. Durbin. "Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignments," *PROTEINS: Structure, Function, and Genetics*, 28, 3, pp. 405-420, 1997.

[15] C. J. A. Sigrist, L. Cerutti, E. Castro, P. S. Langendijk-Genevaux, V. Bulliard, A. Bairoch and N. Hulo. "PROSITE, a protein domain database for functional characterization and annotation," *Nucleic Acids Research*, 38, D161-D166, 2010.

[16] L. C. Freeman. "A set of measures of centrality based on betweenness," *Sociometry*, 40, 1, pp. 35-41, 1977.

[17] L. H. Greene. "Protein structure network", *Briefings in functional genomics*, 2, 6, pp. 469-478, 2012.

[18] R. Jain, H. K. Yalamanchili and N. Parekh. "Identifying Structural Repeats in Proteins using Graph Centrality Measures," In *IEEE proceedings of World Congress on Nature & Biologically Inspired Computing (NaBIC)*, pp. 110-115, 2009.

[19] G. Amitai, A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger and S. Pietrokovski. "Network analysis of protein structures identifies functional residues," *J. Mol. Biol*, 344, pp. 1135-1146, 2004.

[20] S. Vishveshwara, K. V. Brinda, and N. Kannan. "Protein structure: Insights from graph theory," *Journal of Theoretical and Computational Chemistry*, 1, pp. 1, 2002.

[21] S. M. Patra and S. Vishveshwara. "Backbone cluster identification in proteins by a graph theoretical method," *Biophysical Chemistry*, 84, pp. 13-25, 2000.

[22] B. Chakrabarty and N. Parekh. "Analysis of graph centrality measures for identifying ankyrin repeats," In *IEEE proceedings of 2nd World congress on Information and Communication Technologies (WICT)*, pp. 156-161, 2012.

[23] B. H. Junker, D. Koschützki, and F. Schreiber. "Exploration of biological network centralities with CentiBiN," *BMC Bioinformatics*, 7, pp. 219, 2006.

[24] G. Csrdi, and T. Nepusz. "The igraph software package for complex network research," *Inter Journal*, Complex Systems, pp. 1695, 2006.

[25] M. Newman. *Networks: An Introduction*, Oxford University Press, New York, USA, 2010.

Author Biographies

Broto Chakrabarty is a student of MS by Research in Bioinformatics at International Institute of Information Technology, Hyderabad, India. He completed his Bachelor of Technology in Bioinformatics in 2009 from Jaypee University of Information Technology, Wakhnaghat, India. His major area of interest is biological network analysis.

Nita Parekh is an Associate Professor at the Centre for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology, Hyderabad, India. She received her PhD in Physics from Jawaharlal Nehru University, New Delhi, India. Her major areas of interest are Computational Biology and Bioinformatics.