Journal of Theoretical and Computational Chemistry, Vol. 1, No. 1 (2002) 000–000

PROTEIN STRUCTURE: INSIGHTS FROM GRAPH THEORY

SARASWATHI VISHVESHWARA,* K. V. BRINDA and N. KANNAN†
Molecular Biophysics Unit, Indian Institute of Science
Bangalore 560012, India
**sv@mbu.iisc.ernet.in*

The sequence and structure of a large body of proteins are becoming increasingly available. It is desirable to explore mathematical tools for efficient extraction of information from such sources. The principles of graph theory, which was earlier applied in fields such as electrical engineering and computer networks are now being adopted to investigate protein structure, folding, stability, function and dynamics. This review deals with a brief account of relevant graphs and graph theoretic concepts. The concepts of protein graph construction are discussed. The manner in which graphs are analyzed and parameters relevant to protein structure are extracted, are explained. The structural and biological information derived from protein structures using these methods is presented.

Keywords: Graph spectra; eigenvalue; laplacian matrix; topological indices; protein graphs; sub-graph isomorphism; clusters in protein structure; protein folding; non-covalent interactions; protein flexibility.

1. Introduction

Graph Theory is a branch of discrete mathematics, distinguished by the geometric approach to the study of objects. The principal object of the theory is a graph and its generalization. Any problem or object under consideration is represented in the form of nodes (vertices, elements) and edges (connections). Although the topic is more than two centuries old, only in recent times it has gained momentum and has been routinely used in various branches of science and engineering. The mathematics developed earlier can now be applied to systems with large number of vertices and edges, since computers can be effectively made use of in obtaining solutions to such large graphs. Extensive applications of graph theory are made use of in the fields such as electrical circuits, communication and transportation networks.[1] Chemical molecules being a set of atoms or groups of atoms (vertices) connected by covalent bonds (edges) have also been extensively investigated by graph theory.[2] A wealth of information has been derived on electron delocalized molecules by considering electrons and atomic orbitals as vertices and overlap between them as edges.[3–6]

The structure of biopolymers like proteins is governed to a large extent by non-covalent interactions and recently, graph theory is being used to gain insight into the structures of proteins. Non-bonded interactions such as Van der Waal's forces and hydrogen bonds confer unique three-dimensional structures to proteins. Analysis of the topological details of proteins with known structures, such as the clustering of specific types of amino acids important for structure, folding and function, is of great value and is an active field of research. Since the structures of a large number of proteins are being solved by the method of X-ray crystallography, automatic methods of analysis are required to analyze them and recently, the tools from graph theory are being explored for such analysis. Further, the genome research is yielding enormous amount of nucleic acid and protein sequences and the field of proteomics has come into existence, which deals with large scale analysis of proteins including their structural characterization by modeling.

*Corresponding author.
†Current Address: Cold Spring Harbour Laboratory, 1, Bungtown Road, P.O. Box 100, NY, USA.

Protein modeling is another area where developments are taking place and the potentials of graph theory are being explored. The present article aims at reviewing the graph theoretic investigations being carried out on proteins and discusses the future scope of this approach. The presentation is biased towards the clustering algorithm and its applications, due to our involvement with this subject. However, we recognize the contributions of other aspects of graph theory in elucidating the structure, function, folding and dynamics of proteins and have made a brief presentation of these topics as well.

A large number of books and articles dealing with the mathematics of graph theory, its applications and computer algorithms are available.[1,2,7] A brief description of the graphs, properties of the graph which are relevant to the present article is presented in the following section and the formulation of protein structure graphs and their applications are discussed in subsequent sections.

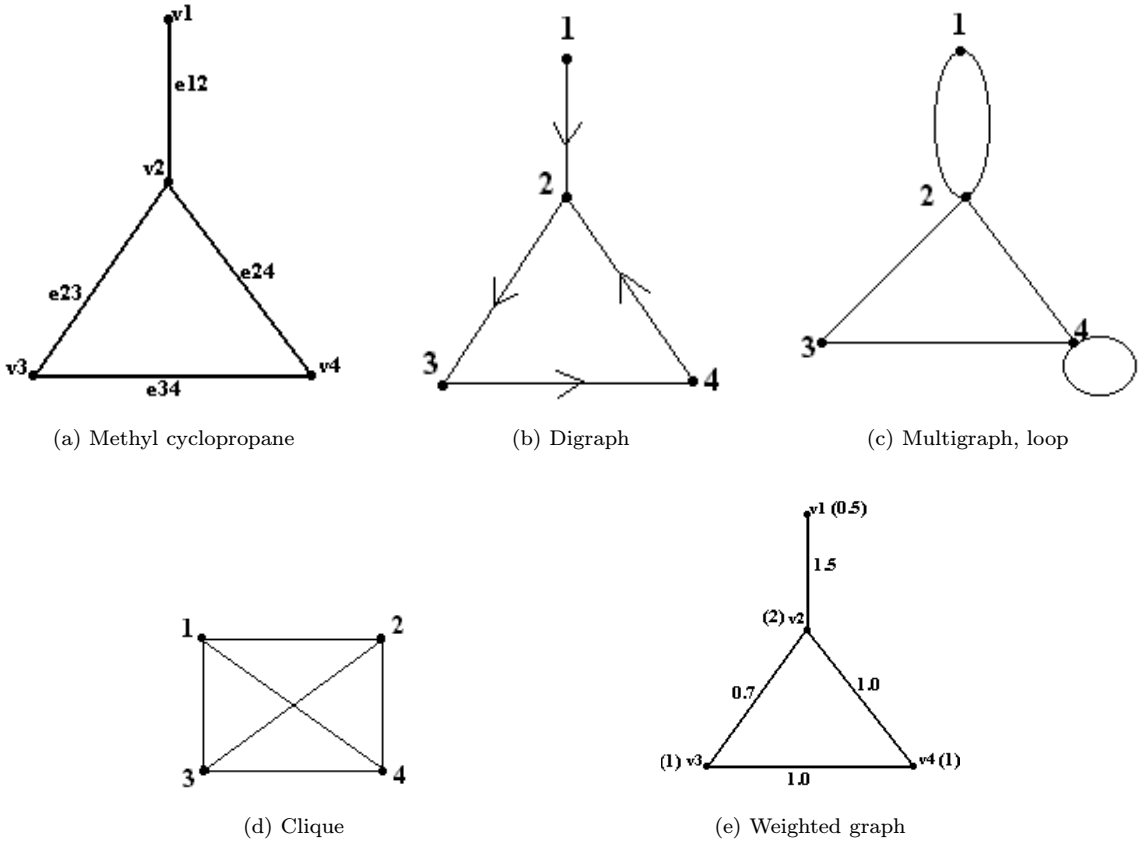
2. Properties of Graphs

Graph theory has been found to be useful in a variety of problems. This has become possible by mathematically representing the graphs and studying its properties and identifying graph invariant parameters. In this section, we deal with the basic concepts of graph theory.

2.1. *Definitions*

2.1.1. *Graph (vertex, edge, degree)*

A graph $G = G(V, E)$ consists of a set of vertices V and a set of edges E, in which the vertices and edges are related as follows. Two vertices v_i and v_j of a graph G are said to be adjacent if there is an edge e_{ij} connecting them. The vertices v_i and v_j are then said to be incident to the edge e_{ij}. Two distinct edges of G are adjacent if they have at least one vertex in common. The degree of a vertex is denoted by $\deg_i$, and



(a) Methyl cyclopropane (b) Digraph (c) Multigraph, loop

(d) Clique (e) Weighted graph

Fig. 1. Types of graphs.

is equal to the number of adjacent vertices to vertex v_i. Graph 1(a) (Fig. 1) consists of 4 vertices and 4 edges, in which the vertex set $V(1) = (v_1, v_2, v_3, v_4)$ and the edge set $E(1) = (e_{12}, e_{23}, e_{34}, e_{24})$ and the degree set $\mathrm{DEG}(1) = (1, 3, 2, 2)$.

2.1.2. *Types of graphs*

A graph can be undirected or directed (digraph) as shown in graphs 1(a) and 1(b) (Fig. 1) respectively. In an undirected graph, the direction of the edges is immaterial or in other words, $e_{ij} = e_{ji}$. Chemical graphs representing bonds between two atoms are represented by undirected graphs. An edge (or edges) in a digraph on the other hand is directed and hence, e_{ij} need not be equal to e_{ji}. The graph representation of a chemical reaction from states A to B is directed. Digraphs are extremely common in electrical circuits, where the current flows in a directed fashion. In this article, we deal with undirected graphs since molecular topology is represented as undirected connection between atoms or groups of atoms.

A multigraph is a graph, in which some of its vertices are connected by more than one edge [graph 1(c)]. The vertices 1 and 2 in graph 1(c) are connected by two edges. A loop is an edge joining a vertex with itself [vertex 4 in graph 1(c)]. Multigraphs are used to represent multiple bonds and loops are used to represent lone pair electrons in case of molecules. Graphs without multiple edges and loops are called simple graphs. Since we are interested in protein topology formed by non-covalent interactions and not on the details of electronic structures, our focus will be on simple graphs.

A complete graph (clique) K_n has every pair of its N vertices adjacent. Graph 1(d) is K_4 and is known as a clique of size 4. Identification of largest clique in a graph helps in graph comparisons, especially when dealing with large graphs.

Weighted graph is a powerful graph representation in which the vertices and edges are discriminated from each other by giving different weights for each of them. The vertices v_1, v_2, v_3 and v_4 in graph 1(e) have weights 0.5, 2, 1 and 1 respectively. The four edges also have varying weights. A chemical structure with hetero atoms and unequal bond lengths are represented by an edge and a vertex weighted graph.

2.1.3. *Subgraph*

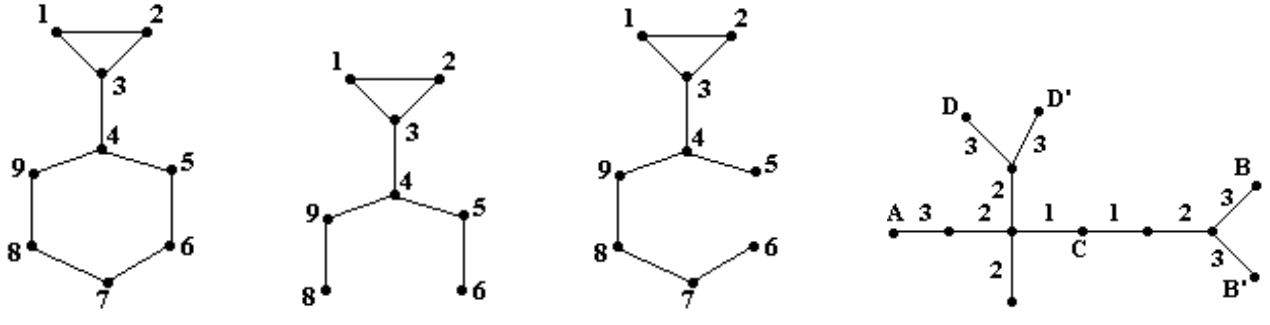
A subgraph G' of G is a graph whose vertices and edges are contained in G. The graphs 2(b) and 2(c) (Fig. 2) are subgraphs of 2(a). The subgraph $(G - v_i)$ is obtained from the graph G by deletion of the vertex v_i and its incident edges. The vertex v7 in graph 2(a) is deleted to obtain the graph 2(b). The subgraph $(G - e_{ij})$ is obtained from graph G by deletion of the edge e_{ij}. The graph 2(c) is obtained from graph 2(a) by deleting the edge e56.

2.1.4. *Center of a graph*

The distance between two vertices (separation or path length) in a graph is measured in terms of the number of edges connecting the two vertices. Eccentricity,[1] $E(v)$ of a vertex v in a graph G is the distance from v to the vertex farthest from v in G, i.e.

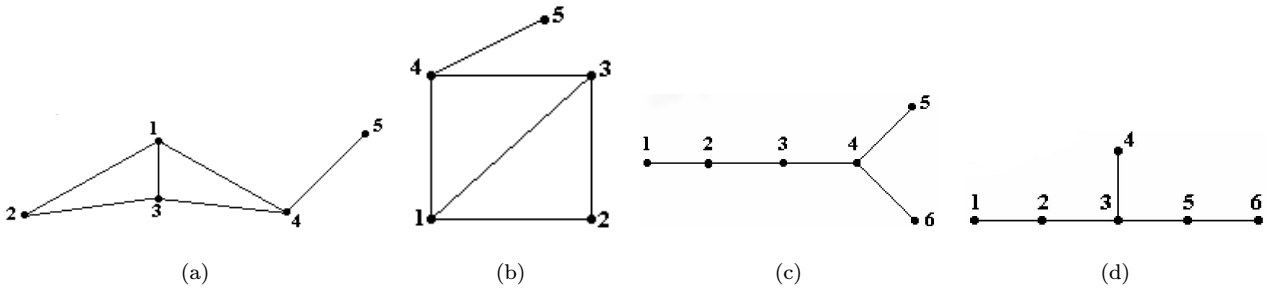
$$E(v) = \max_{v_i \in G} d(v, v_i)\,. \tag{1}$$

Now the vertex with minimum eccentricity is defined as the center of the graph.[1,8] The maximum



(a) (b) (c) (d)

Fig. 2. Subgraphs and the center of graph.



(a) (b) (c) (d)

Fig. 3. Graph isomorphism.

eccentricity in graph 2(d) is 6, which is taken by vertices such as A, B and D. The minimum eccentricity (3) is taken by the vertex C and is considered as the center of the graph.

2.1.5. *Graph isomorphism*

Two graphs G and G' are said to be isomorphic to each other if there is one-to-one correspondence between their vertices and between their edges such that the edge–vertex relationship is preserved. Graphs 3(a) and 3(b) (Fig. 3) are isomorphic graphs. Two isomorphic graphs must have the same number of vertices, same number of edges and equal number of vertices with a given degree. However, these conditions are by no means sufficient to declare that two graphs are isomorphic and a number of algorithms are available to detect graph isomorphism.[1] For example, graphs 3(c) and 3(d) have the same number of vertices, same number of edges and equal number of vertices with a given degree. But they are not isomorphic graphs because their connectivities are different as can be seen by visual inspection.

2.2. *Matrix representation*

Pictorial graphs are useful in visualization. However, a graph can be converted into an algebraic form of matrix. When a graph is represented in a matrix form, it becomes a mathematical entity on which operations can be performed. Analytical solutions can be obtained and numerical algorithms can be applied. Solutions to large graphs can be obtained by standard computer algorithms. The reason for renewed interest in graph theory is because complicated networks in various branches of science and engineering such as electrical engineering, circuit netlist, computer networks, large chemical and biological molecules can be represented as matrices and solutions can be easily obtained by computer algorithms. Graph comparisons, quantitative characterizations, computation of topological indices, clustering and partitioning are some of the major computations which have yielded valuable results in various disciplines. The type of matrix representation depends on the property that one is looking for. The most common one is Adjacency matrix, which contains the basic information on graph connectivity.

2.2.1. *Adjacency matrix*

The Adjacency matrix $A = A(G)$ of an undirected graph G with N vertices is the square $N \times N$ symmetric matrix whose ijth elements are defined as:

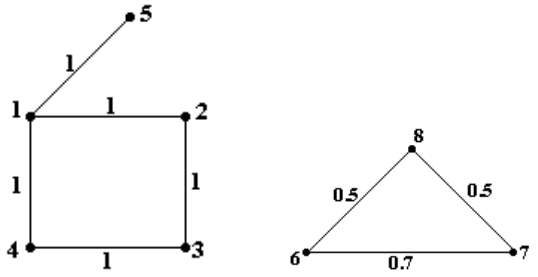
$$[A]_{ij} = W_{ij} \text{ if } i \neq j \text{ and } i \text{ and } j \text{ are adjacent vertices (i.e. connected)}$$

$$0 \text{ if } i = j \text{ or there is no edge between } i \text{ and } j$$

$$W_{ij} = 1, \text{ for an unweighted graph and takes the value of the weight of the edge } e_{ij}, \text{ in a weighted graph.}$$

The unweighted Adjacency matrix for methyl cyclobutane [graph 4(a), Fig. 4] is given as:

$$A = \begin{matrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{matrix}$$



(a) (b)

Fig. 4. Graph example for matrix representation.

2.2.2. *Laplacian matrix*

Laplacian matrix is also known as Kirchoff matrix or admittance matrix because of its association with Kirchoff's theorem and conductivity in electrical connections. The Laplacian matrix of a graph G, $L = L(G)$, is defined as:

$$L(G) = \mathrm{DEG}(G) - A(G)(\text{or } L = D - A) \qquad (2)$$

where $\mathrm{DEG}(G)$ is the degree matrix and $A(G)$ is the Adjacency matrix of the graph G.

The degree matrix is a diagonal matrix, which has the information regarding the degree of each vertex. It is obtained by summing up each of the columns (or rows) of the Adjacency matrix.

The degree matrix for graph 4(a) is given as:

$$D = \begin{matrix} 3 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{matrix}$$

and the Laplacian matrix for graph 4(a) is:

$$L = \begin{matrix} 3 & -1 & 0 & -1 & -1 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ -1 & 0 & -1 & 2 & 0 \\ -1 & 0 & 0 & 0 & 1 \end{matrix} .$$

2.3. *Graph spectra*

Graph spectral theory is concerned with the relationships between the algebraic properties of the spectra of certain matrices such as the Adjacency or Laplacian matrices associated with a graph and the topological properties of the graph. The eigenvalues and eigenvectors of matrices associated with a graph are the most important graph spectral parameters, which provide information on the structure and topology of the graph and analysis of these quantities is known as graph spectral analysis. Many of the mathematical proofs, which relate the structure of the graph to its spectra, is given in the books by Norman Biggs[9] and Strang.[10] Spectral techniques are used most often in the design of circuits, VLSI chips and computer networks. Identification of clusters and similarity in connectivity patterns can be deduced from the spectral analysis of connected graphs.[11,12] Graph spectra is also extensively used in chemical graph theory to derive topological indices such as the resonance energy, molecular orbital energy and topology of π electron systems.[2,13,14] Graph indices such as Weiner index,[15,16] Hosoya index[17] and indices derived from graph spectral analysis are extensively used in quantitative structure-activity relationship (QSAR) studies.[3,5,18,19] Configuration statistics such as the distribution function of the radius of gyration, inertial tensor and partition function of polymer chains have been derived from the spectra of Laplacian matrix.[20,21] The graph spectral analysis has yielded valuable results in the identification of clusters in protein structures[22,23] and this part will be dealt in great detail in a later section.

The eigenvalues and vector components of the Adjacency and Laplacian matrices of graph 4(a) given in Tables 1(a) and 1(b) respectively, give us significant information about the graph. The vector components corresponding to the largest eigenvalue contains the information regarding the contribution of each node to the graph. In graph 4(a), node 1 has 3 edges, nodes 2, 3 and 4 have 2 edges each and node 5 has only one edge. The magnitude of the vector components of the largest eigenvalue of both the matrices reflects this observation. Nodes 2 and 4 have degenerate vector components because they are identical in nature.

3. Graph Spectral Properties Relevant to the Article

3.1. *Spectra of isomorphic graphs*

As mentioned earlier, there is no unique way of detecting graph isomorphism. Graph spectral analysis

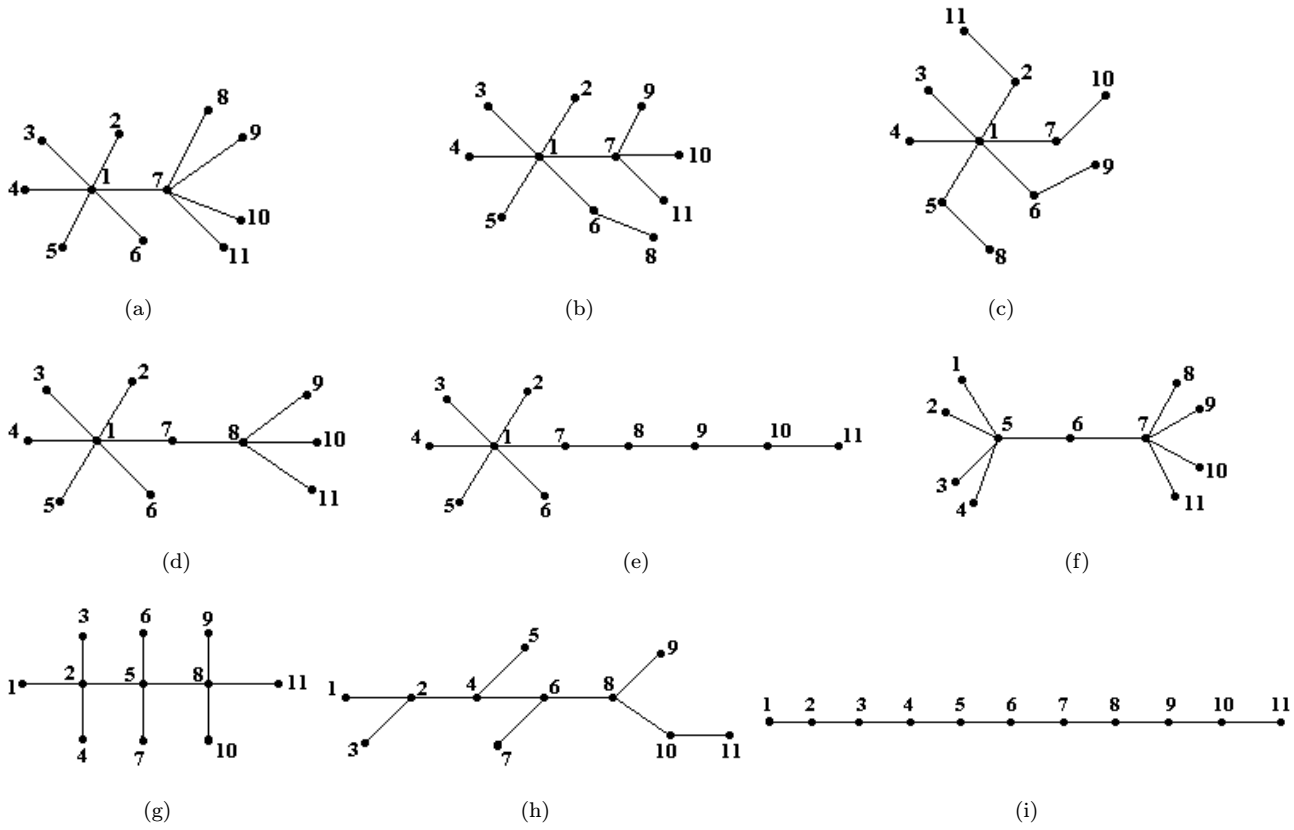


Fig. 5. Graph with 11 nodes and 10 vertices (Examples chosen to elucidate eigen spectra).

Table 1(a). Spectra obtained from Adjacency matrix of methyl cyclobutane shown in graph 4(a).

Eigne Values	−2.1358	−0.6622	0.0000	0.6622	2.1358
Node	EVC[1]	EVC	EVC	EVC	EVC
1	−0.5573	−0.4352	0.0000	0.4352	**−0.5573**[2]
2	0.4647	−0.1845	0.7071	−0.1845	−0.4647
3	−0.4352	0.5573	−0.0000	−0.5573	−0.4352
4	0.4647	−0.1845	−0.7071	−0.1845	−0.4647
5	0.2610	0.6572	0.0000	0.6572	−0.2610

Table 1(b). Spectra obtained from Laplacian matrix of methyl cyclobutane shown in graph 4(a).

Eigen Values	−0.0000	0.8299	2.0000	2.6889	4.4812
Nodes	EVC	EVC	EVC	EVC	EVC
1	0.4472	0.1380	0.0000	0.5362	**0.7024**
2	0.4472	−0.2560	−0.7071	0.2422	−0.4193
3	0.4472	−0.4375	0.0000	−0.7031	0.3380
4	0.4472	−0.2560	0.7071	0.2422	−0.4193
5	0.4472	0.8115	0.0000	−0.3175	−0.2018

[1]Eigenvector components.

[2]The largest vector component of the largest eigenvalue [lvc(L) is shown in bold.

Table 2. Eigen spectra of Adjacency matrix of graphs in Fig. 5.

Graph	5(a)	5(b)	5(c)	5(d)	5(e)	5(f)	5(g)	5(h)	5(i)
lev[1]	2.690	2.633	2.589	2.533	2.500	2.450	2.450	2.232	1.932
Node	vc(L)[2]	vc(L)	vc(L)	vc(L)	vc(L)	vc(L)	vc(L)	vc(L)	vc(L)
11	0.186	0.163	0.117	0.107	0.016	0.204	0.183	0.101	0.106
10	0.186	0.163	0.117	0.107	0.040	0.204	0.183	0.226	0.204
9	0.186	0.163	0.117	0.107	0.084	0.204	0.183	0.181	0.289
8	0.186	0.108	0.117	0.271	0.171	0.204	0.447	0.403	0.354
7	0.500	0.428	0.303	0.365	0.342	**0.500**	0.224	0.221	0.394
6	0.224	0.284	0.303	0.258	0.274	0.408	0.224	**0.493**	**0.408**
5	0.224	0.243	0.303	0.258	0.274	**0.500**	**0.548**	0.214	0.394
4	0.224	0.243	0.258	0.258	0.274	0.204	0.183	0.477	0.354
3	0.224	0.243	0.258	0.258	0.274	0.204	0.183	0.160	0.289
2	0.224	0.243	0.303	0.258	0.274	0.204	0.447	0.357	0.204
1	**0.602**[2]	**0.640**	**0.667**	**0.653**	**0.685**	0.204	0.183	0.160	0.106

[1]Largest eigenvalue.
[2]Vector components corresponding to largest eigenvalue.
[3]lvc(L) are shown in bold.

gives information on isomorphism. From the spectral point of view, isomorphic graphs have the same spectra, and hence they are said to be cospectral. However, the converse need not be true. Finding simple criteria and efficient computer algorithms to detect isomorphic graphs is an active area of research.[24] Detection of graph or subgraph (part of a graph) isomorphism is a powerful technique in protein structure comparison.[25–27]

3.2. *Largest eigenvalue (lev)*

The largest eigenvalue (lev) depends upon the highest degree in the graph. For any k regular graph G (a graph with k degree on all the vertices), the eigenvalue with the largest absolute value is k. A corollary to this theorem is that the lev of a clique of n vertices is $n-1$. In a general connected graph, the lev is always $\leq$ to the largest degree in the graph. In a graph with n vertices, the absolute value of lev decreases as the degree of vertices decreases. The lev of a clique with 11 vertices is 10 and that of a linear chain with 11 vertices [graph 5(i) in Fig. 5] is 1.932 and those graphs with 11 nodes with varying degrees in between, take up lev within the limits of these two. So, the upper bound for lev is $(n-1)$, where n is the number of vertices.

In an irregular graph, it is difficult to analytically deduce lev. The lev in between the two bounds depends on the nature of connectivity. This is illustrated with examples of 11 nodes and 10 edges systems given in graphs 5(a)–5(i) (Fig. 5) and their lev and the corresponding vector components are given in Table 2. In graphs 5(a)–5(e), the highest degree is 6. In graphs 5(f)–5(i), the highest degree is 5, 4, 3 and 2 respectively. It can be noticed that the lev is generally higher if the graph contains vertices of high degree. The lev decreases gradually from the graph with highest degree 6 to the one with highest degree 2. In case of graphs 5(a)–5(e), where there is one common vertex with degree 6 (highest degree) and the degrees on the other vertices are different (less than 6 in all cases), the lev also depends on the degree of the vertices adjoining the highest degree vertex.

3.3. *Eigenvector components and branching of nodes*

Let us consider a graph G with "n" vertices. Let $x_1, x_2, \ldots, x_n$ be the weights assigned to "n" vertices. We try to find a numerical value to x_i, which is proportional to the sum s_i, of all the edges emanating from vertex "i" in the graph G. In other words, if there are "k" vertices connected to the vertex "i",

then x_k are to satisfy, in a non-trivial way, the system of homogeneous linear equations

$$\lambda x_i = \frac{1}{d_i} \sum_{k \cdot i} x_k \tag{3}$$

the value of λ being suitably chosen and d_i being degree of vertex "i". The above equation can be written as:

$$\lambda DX = AX \tag{4}$$

where A corresponds to the Adjacency matrix and D the diagonal matrix of G and X denotes the corresponding eigenvector with components x_k. Thus, the eigenvector components may be directly interpreted as "weights" of the corresponding vertices in the graph. The weights are direct indication of the extent of connectivity of the corresponding vertices in the graph. This concept is explained using some simple examples below.

Let us consider the graphs 5(a)–5(i) shown in Fig. 5. The lev and the corresponding vector component values [vc(L)] of the Adjacency matrices of the graphs 5(a)–5(i) are given in Table 2. It can be noticed that the largest vector component (lvc) of the lev [lvc(L)] corresponds to either the node with highest degree or to the center of the graph. The first preference for the lvc(L) is the node with the highest degree. For example, in graphs 5(a)–5(e), node 1, which has the highest degree, has the lvc(L). In graph 5(f), nodes 5 and 7 have degree 5, which is the highest degree in the graph and both of them have degenerate values of vc(L). However, in case there are many nodes with degenerate highest degree, the one which is closest to the geometric center of the graph, takes up the lvc(L). This can be seen in graphs 5(g)–5(i), where the center of the graph has taken the lvc(L). It can also be noted that the removal of the vertex with lvc(L) results in a subgraph, which is smaller than the one that can be obtained by removing any other vertex. Similarly, removal of a vertex with the smallest vector component corresponding to the lev [svc(L)] results in a subgraph, which is larger than the one obtained by removal of any other vertex. The contributions of the nodes in the graph, decreases as we move away from the lvc(L) and this is reflected in the magnitude of the vector components corresponding to the lev. In summary, the vector components of a graph can be interpreted as a direct measure of the contribution of each vertex to the overall connectivity of the graph.

The eigen spectral concept has been used in Huckel molecular orbital theory to obtain the resonance energy levels and the composition of the corresponding molecular orbitals.[3–6] The π electron atomic orbitals are considered as nodes and edges are made between interacting orbitals. Weights are assigned to represent the energy of interaction. The resulting eigenvalues of the molecular graph are correlated with the energy of the molecular orbital and the corresponding vector components represent the contribution of the atomic orbital to the concerned molecular orbital. The spectral properties are made use in identifying important nodes, which contribute to the stability of clusters in protein structures (such as cluster centers), and will be discussed in a later section.

3.4. *Laplacian matrix and graph spectra*

The Laplacian matrix $L(D - A)$ of a graph contains the specific degree information on each of the vertex as elements of the diagonal. Some of the properties of the Laplacian matrix are given below. In the properties shown below X denotes the eigenvector of the Laplacian matrix and x_i denotes the vector components.

(a) The Laplacian matrix is symmetric and non-negative definite, i.e.

$$\left(X^T L X = \sum_{ij} L_{ij} x_i x_j \geq 0 \right) \tag{5}$$

and all the eigenvalues of L are ≤ 0.

(b) The smallest eigenvalue of L is 0 with eigenvector

$$X = \left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}} \right) \tag{6}$$

where n is number of vertices in the graph.

(c) The inner product $X^T L X$ can be written as

$$\begin{aligned} &= X^T D X - X^T A X \\ &= \sum_i d_i x_i^2 - \sum_{i,j} A_{ij} x_i x_j \\ &= \sum_i d_i x_i^2 - 2 \sum_{i,j} x_i x_j \end{aligned} \tag{7}$$

which can be written as

$$X^T L X = \sum_{i,j} (x_i - x_j)^2 \,. \tag{8}$$

3.5. *Relation between Adjacency and Laplacian matrices*

It is clear from the relation $L = D - A$ that the Laplacian and the Adjacency matrices are related in a very simple way. For graphs, which are regular with degree k, the eigenvalues of the Adjacency have a linear relation with the eigenvalues of the Laplacian. For example, if

$$0 = \theta_1 \leq \theta_2 \cdots \leq \theta_n$$

be the eigenvalues of the Laplacian in the increasing order. Let

$$\lambda_1, \lambda_2 \cdots \lambda_n$$

be the eigenvalues of the Adjacency matrix, then a relation of the form

$$\theta_i = k - \lambda_i \tag{9}$$

can be written between the eigenvalues of the Laplacian and the Adjacency. This is quite straightforward for regular graphs as the Laplacian matrix for k regular graphs can be written as

$$L = kI - A\,. \tag{10}$$

This linear relation does not exist for graphs, which are non-regular. Since most of the graphs of our interest are non-regular, one cannot define a clear relation between the spectra of the Adjacency and the Laplacian.[28]

3.6. *Clustering by graph spectra*

Identifying clusters is an important operation carried out in the field of electrical network connections.[29–32] There are several algorithms to do this operation by traditional matrix manipulations.[1,33] The results, however, depend sometimes on the method adopted and it requires several iterative steps to achieve clustering. Further, such operations can be performed only on matrices with binary values of zero and one. Graph spectral method is a powerful tool, which can yield unique results by a single numeric computation.[29] Further, it can also be used to get clustering information on weighted graphs. These concepts are adopted to obtain non-bonded clusters in protein structures[22,23] and the following algorithm is used for such investigations.

An Adjacency matrix of a weighted graph can be constructed by assigning weights to the edges connecting the vertices. Now the clustering problem is to find the location of "n" vertices which minimizes the weighted sum of squared distances between the points which is given by the function[34]:

$$Z = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 A_{ij}\,. \tag{11}$$

The details of clustering procedure[34] is given in Appendix A. The essential point of interest is that the second smallest eigenvalue of the Laplacian matrix and its associated vector components yield the clustering of points in the graph. This is illustrated with an example. Consider the Laplacian matrix for graph 4(a) (Fig. 4) presented in Sec. 2.2.2. We combine graph 4(a) and graph 4(b) and construct a Laplacian matrix with edge weights $(1/d_{ij})$, where d_{ij} is the distance between vertices i and j. The distances between the vertices of graph 4(a) and graph 4(b) are considered to be very large (say 100) and thus the matrix elements corresponding to a vertex from graph 4(a) and the other from graph 4(b) is considered to have a very small value of 0.01. The Laplacian matrix of 8 vertices thus considered is diagonalized and their eigenvalues and corresponding vector components are given in Table 3. The vector components corresponding to the smallest eigenvalue yields a trivial solution as all values are degenerate. The vector components corresponding to the second smallest eigenvalue contains the desired information about clustering, where the cluster forming residues have identical values. In Fig. 4, nodes 1–5 form a cluster (cluster 1) and 6–8 form another cluster (cluster 2). This information can be obtained by inspecting the vector component values of the second smallest eigenvalue because nodes 1–5 have the same value and nodes 5–8 have the same value. Thus, clusters can be identified from the vector components of the second smallest eigenvalue. Additionally, the larger eigenvalues contain information regarding only one of the clusters. If we look at the vector components of the largest eigenvalues corresponding to each of the two clusters, we can see that the node with the largest vector component is the one with the highest degree. In case of cluster 1, node 1 has the highest degree and hence it has the largest vector component. And in cluster 2, nodes 6 and 7,

both have the same degree and identical weights on their edges. Hence, they are equivalent and therefore, have degenerate vector component values. Though node 8 has the same degree as nodes 6 and 7, it is different from nodes 6 and 7 because of the edge weights. The nodes with the highest degree and highest edge weights have the largest vector components in the top eigenvalues.

4. Protein Graphs

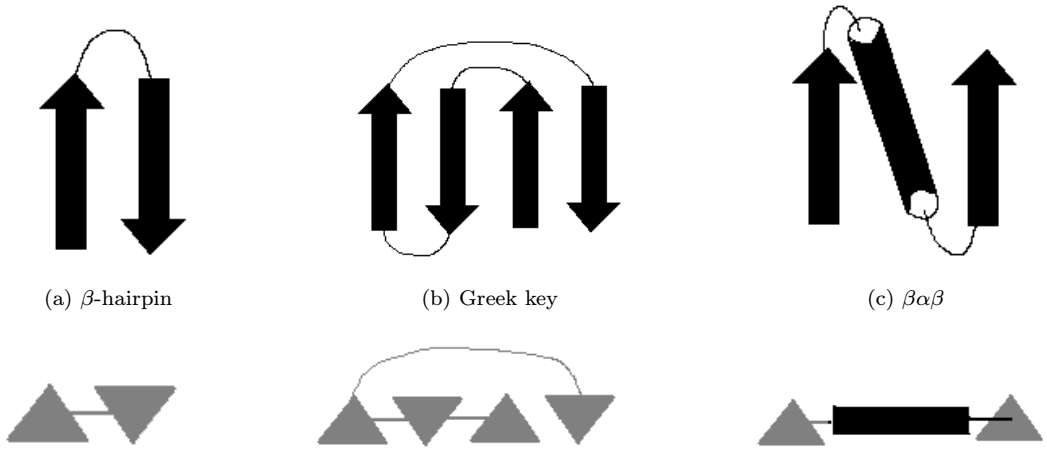
The three dimensional structure of proteins is the key to understanding their function and evolution. It is often stated that the folding of proteins to its unique native state is the second genetic code. Analysis of stable folded three-dimensional structures of proteins provides insights into their folding stability and

Table 3. Eigen spectra of the Laplacian matrix of the combined graphs 4(a) and 4(b) (Fig. 4).

Eigen Values	0.0000	0.0800	0.9016	1.5500	1.9500	2.0600	2.7420	4.5164
Node	EVC[1]	EVC	EVC	EVC	EVC	EVC	EVC	EVC
1	0.3536	0.2739	0.1380	−0.0000	0.0000	0.0000	0.5362	**0.7024**[2]
2	0.3536	0.2739	−0.2560	−0.0000	−0.0000	0.7071	0.2422	−0.4193
3	0.3536	0.2739	−0.4375	−0.0000	−0.0000	−0.0000	−0.7031	0.3380
4	0.3536	0.2739	−0.2560	−0.0000	−0.0000	−0.7071	0.2422	−0.4193
5	0.3536	0.2739	0.8115	0.0000	−0.0000	−0.0000	−0.3175	−0.2018
6	0.3536	−0.4565	−0.0000	−0.4082	−0.7071	0.0000	0.0000	0.0000
7	0.3536	−0.4564	0.0000	−0.4082	0.7071	−0.0000	−0.0000	−0.0000
8	0.3536	−0.4564	−0.0000	0.8165	−0.0000	0.0000	0.0000	0.0000

[1]Eigen vector components.

[2]lvc(L) is shown in bold.



(a) β-hairpin (b) Greek key (c) $\beta\alpha\beta$

(d) Graph representation of β-hairpin (e) Graph representation Greek key (f) Graph representation of $\beta\alpha\beta$

Fig. 6. Some of the common motifs in proteins (a, b and c) and their graph representations (d, e and f). The β-strand are represented as arrows pointing from the N-terminal towards and C-terminal and the α-helix is represented as a cylinder. The connectivities between these secondary structures are shown as loops in the motifs. The secondary structures are the nodes (β-strand is represented as triangle and α-helix as a rectangle) and their spatial connectivites are the edges in the graph representations (d, e and f).

function. Results of analysis have also become foundation for protein structure prediction from amino acid sequence, design of novel proteins with desirable catalytic function and rational drug design studies.

Properties of graphs and their spectra discussed in the previous sections are made use of in elucidating protein structures. The main problem is to define nodes and edges. Depending on what is defined as a node and an edge, different aspects of protein structures can be characterized. Sub-graph isomorphism,[25–27] clustering,[22,23] correlation of eigen spectra with topological and physical properties[35,36] are some of the topics that are addressed in literature.

4.1. *Protein structural topology*

A protein structure has geometry, expressed in the conformation of the protein backbone and side-chains. Many structures can differ in terms of the conformational features (geometry) but still can have the same topology (the gross shape). Over many years, many authors have studied protein structural topology and a number of definitions and notations have emerged. The early work related to structural topology was through hand drawn diagrammatic form.[37–39] These diagrams were two-dimensional schematic representations of protein folds with particular symbols used to represent helices and strands. Figures 6(a)–6(c) show diagrammatic representations of a few common motifs like β-hairpin, Greek key and $\beta\alpha\beta$. This kind of representation helped in understanding how particular folding pathways would favor particular types of topology and to what extent topological similarities between structures might imply an evolutionary relationship.

4.2. *Mathematical formulation of protein topology*

A more mathematical formulation of protein structural topology was made by Koch and co-workers,[40] wherein they introduced the concept of a mathematical graph to represent β structures. In a β-graph, the vertex represents a single β-strand and the two edge sets describe sequential and hydrogen bond connections respectively. The language of graph theory helped in representation of any topology however complicated. Later Grigoriev and co-workers[41] represented all α-helical structures in the form of connected graphs. Again the approach uses graph theoretical techniques in which the nodes represent secondary structures and the edges represent contacts between helices, rather than hydrogen bonds which was the case with β-graphs. Several definitions of helical contacts were tested, which resulted in number of graphs with different connectivites. The main application of such an approach was to gain insights into the folding process and domain organization of the proteins. This method of representing the secondary structures as nodes and their connectivities as edges is extensively used for protein structure comparison.[25–27,42] The side chains/main chains and their connectivities are used as a nodes and edges in the recognition of structurally and functionally important patterns and in protein modeling studies. The protein connectivity as whole, irrespective of the nature of secondary structure, is captured by identifying the main chain atoms that are spatially close within a

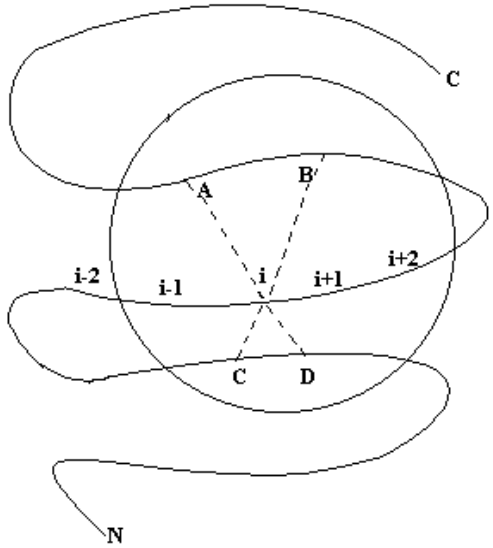


Fig. 7. A schematic representation of the sequential and spatial neighbors of a residue "i" in a polypeptide chain. A distance cut off is taken and the residues which fall within this radius are the spatial neighbors of residue "i". Residues A, B, C and D are spatial neighbors of "i" whereas $i-1$, $i-2$, $i+1$ and $i+2$ are sequential neighbors of i. The spatial neighbors are indicated using dotted lines and the sequential neighbors are shown as continuous segment of the polypeptide chain.

Table 4. Protein graph description.

S. No	Nodes	Edges	Graph Operation	Purpose	References
1	Secondary structure (α-helix, β-strand)	Spatially close Secondary structures	Identification of subgraph isomorphism	Fold and pattern identification	Mitchell *et al.* 1989[25]; Grindley *et al.*, 1993[26]
2	Secondary structure (α-helix, β-strand)	Spatially close Secondary structures (dynamically arrived at)	Dynamical matrix construction	Testing folding rules	Przytycka *et al.*, 2002[42]
3	Side chain	Spatial proximity	Identification of subgraph isomorphism	Functionally and structurally important motif recognition	Artymiuk *et al.*, 1994[27]
4	Side chain	Spatial proximity decided by overlap cut off criterion (weighted edge)	Graph spectra, identification of clusters and cluster centers	Identification of clusters important for function, structure and folding	Kannan and Vishveshwara, 1999[22]
5	Backbone	Spatial neighbours within radius cut off ($6.5 - 7.0$ Å)	Graph spectra, identification of clusters and cluster centers	Identification of proteins with similar folds	Patra and Vishveshwara., 2000[23]
6	Backbone	Spatial neighbours within radius cut off (7.0 Å)	Graph spectra	Protein dynamics	Bahar, 1999[35]
7	All atoms	Defined based on constraints (weighted edge)	Graph spectra	Protein dynamics	Jacobs *et al.*, 2001[36]

prescribed distance (around 7Å, which represent the radial distribution.[43,44] An example of the edges which result from such a representation is shown in Fig. 7 and is made use of in identifying backbone clusters[23] and in GNM model[35] (models to extract dynamical information from protein static structure).

Nodes and edges in protein structures as defined by various groups and the motivation for constructing such graphs is given in Table 4. The basic unit of a protein is its amino acid residue. This is used as the node and the three-dimensional connectivities between these amino acid residues (backbone and/or side chain) is taken as edges for graph construction in cluster identification studies.[22,23] On a different level, the secondary structures (helices and strands) are considered as nodes for pattern identification studies.[25,26] Secondary structure motifs like β-hairpin [Fig. 6(a)] and Greek key [Fig. 6(b)] motifs are considered as nodes [Figs. 6(d) and 6(e)] to study the rules of protein folding.[42] The method of defining nodes and edges for identifying side chain clusters in protein structures[22] is described in detail in Appendix B.

5. Applications of Graph Theoretic Concepts for Protein Structure Analysis

So far we have examined the protein structures and learnt to represent them in the form of graphs and matrices. In this section, we present the important results obtained on protein graphs. A range of exciting and important problems related to protein structure is addressed, which includes recognition of patterns such as protein topology (folds) and active-site motifs, detection of a variety of functionally and

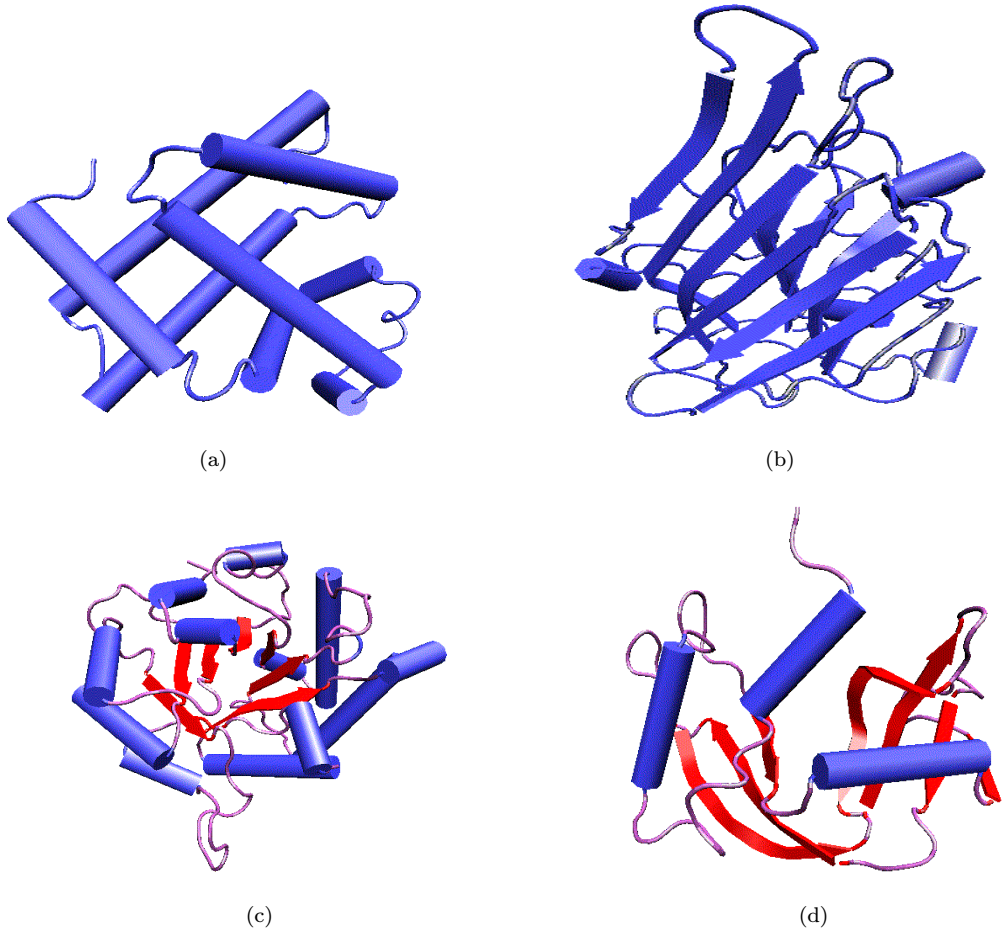
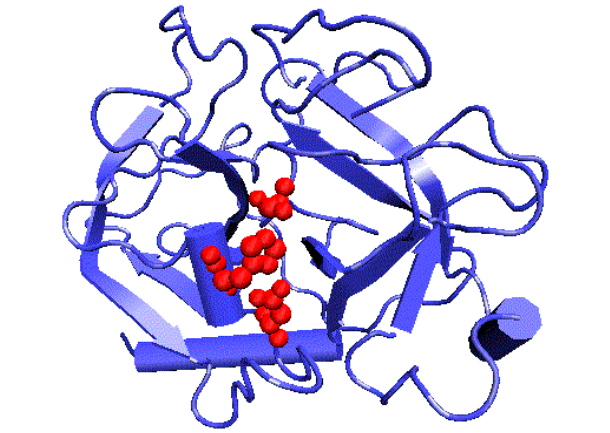


Fig. 8. Protein folds and their representations.[96] The helices are shown as cylinders and strands as shaded arrows. Different protein folds representing all the major protein classes, namely the all α-class (myoglobin), all β-class (immunoglobulin), α/β-class (triose phosphate isomerase) and $\alpha+\beta$ class (Ribonuclease A) are shown in a, b, c and d respectively. It can be seen that the all-α class consists mainly of α-helices, the all β-class is mainly made up of β-strands, the α/β class consists of parallel β-sheets connected through α-helices and the $\alpha+\beta$ class has both α-helices and β-strands, but they do not occur alternatively as in α/β class.

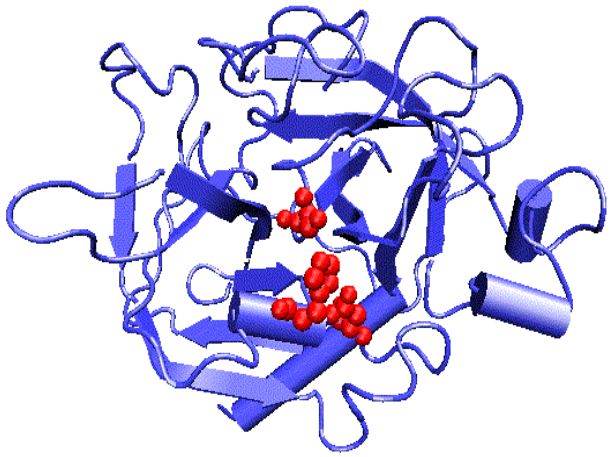
structurally important amino acids, protein structure prediction and extracting dynamical information from static structures.

5.1. *Protein structure comparison: Pattern identification in proteins by graph isomorphism*

It is known that although the number of proteins in nature is enormous, they adopt a limited number of three-dimensional structures,[45] which are represented as folds, families and superfamilies and motifs.[46] A few samples of such protein folds are presented in Figs. 8(a)–8(d). One of the present challenges is to identify the fold adopted by the polypeptide chain and to sort out similarities in protein structures.[47] Further, the proteins also take up specific side chain patterns to carry out their biological functions and identification of such motifs is extremely helpful in structure-function correlation. An example of the catalytic unit of a class of proteins known as serine proteases is given in Fig. 9. Both trypsin [Fig. 9(a)] and elastin [Fig. 9(b)] are serine proteases and have the characteristic catalytic triad consisting of



(a)



(b)

Fig. 9. Active site cluster in serine proteases. The active site cluster consisting of Ser-His-Asp triad in trypsin and elastin are shown using Van der Waal's spheres in a and b respectively. This catalytic triad pattern is characteristic of the serine proteases and is known to be conserved.

Ser-His-Asp. Graph-theoretic approaches have been successfully used in such pattern identification studies. The basic task again is to specify what nodes and edges are. Once a graph is constructed, the patterns can be identified by recognizing different levels of isomorphism.

The process of node and edge identification begins from the co-ordinates of protein atoms obtained from Protein Data Bank.[48] The regular secondary structures, like α-helices and β-strands can be detected from crystallographic co-ordinates using parameters such as hydrogen bonds[49] or the dihedral angles ϕ or ψ.[50] The major axes of helices and strands and their direction can then be evaluated from the co-ordinates of selected atoms. The identified secondary structures are considered as labeled nodes. The distances and angles between the secondary structures are then evaluated. Based on selection criteria, edges are made between nodes and a protein secondary structure graph is constructed.[25,26] An example of such a secondary structure graph is given in Figs. 6(d)–6(f). The helices and strands occurring in proteins are represented as nodes and their three-dimensional connectivities considered for edge formation. On the other hand, the three dimensional patterns of side chains require a different definition of nodes and edges. A side chain is defined by two pseudoatoms, one near the start (S) and the other near the end (E) of the functional part of the side chain.[27] The distances and angles between the pseudoatoms of functionally important side chains are considered as edges. The nodes and edges thus identified in a protein with a well-known pattern are taken as the query. Such patterns are searched in proteins with unrecognized patterns where the edges are defined with some tolerance criteria. Obviously, stringent criteria can pick up patterns, which are very close to the query and relaxing them can broadly recognize the patterns.

Two protein graphs can be compared to see if they share common features by graph isomorphism detection methods. A number of methods are available for subgraph isomorphism detection.[6] Ullman's algorithm[24] has been extensively used by Willet's group. It is a tree searching algorithm which compares successive subgraph isomorphism from a query motif with the protein structure. Matrices are set up for the query and a selected protein structure and the subgraph isomorphism is detected by a series of matrix permutations. The method is further refined to identify maximal common subgraph (MCS). An MCS algorithm allows one to determine the largest subgraph that is common to a pair of graphs. This approach can highlight areas of structural overlap, which may involve structural/functional commonalities, which are not obvious from other methods. The problem of MCS identification however is computationally demanding since it involves up to $P!Q!/R!(P-R)!(Q-R)!$ comparisons, where the two graphs contain P and Q nodes, with R nodes in common.[51] A number of

heuristic methods are available,[51–53] which cleverly reduce the number of comparisons. Since measurement of protein structural similarity using contact map overlap and sub-graph isomorphism is computationally expensive, approximation algorithms have been used for this purpose.[54] The clique detection algorithm has been found to be highly efficient in identifying MCS[55,56] and this procedure is extensively used to detect tertiary structural resemblance in proteins.[26]

The patterns characteristic to a number of proteins such as serine protease, staphylococcal nuclease, NAD binding motif, calcium binding motif etc., are used as query to search across PDB.[48] For instance, serine proteases are characterized by the catalytic triad (ser-his-asp). Figures 9(a) and 9(b) show this catalytic triad in two serine proteases, trypsin and elastin. The pattern is recognized not only in serine proteases, but an additional triad is also found in their precursor proteins (known as zymogens) at a second site. Such unexpected hits prompt one to search for their functions.[27] Subgraph isomorphism can also aid in correlating protein topology or structure with function. Striking resemblances are detected between proteins with low sequence similarity such as carboxy peptidase and leucine aminopeptidase,[57] ribonuclease *H* and connection domains of HIV reverse transcriptase,[58] bacterial signal transduction proteins and *G* proteins,[59] by subgraph isomorphism method.

5.2. *Clusters in protein folding and function*

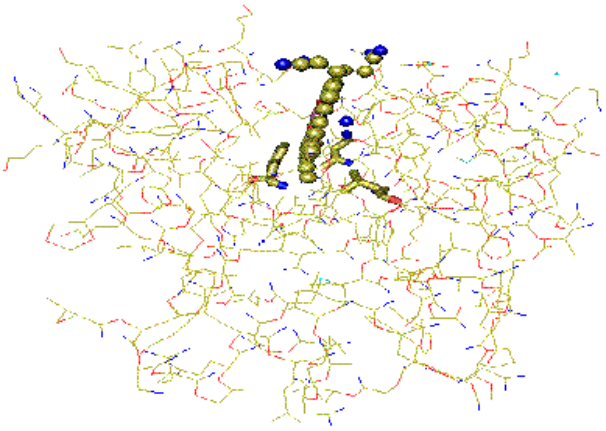
The interaction among residues, together with their interactions with the surrounding, determines the native functional structure of the protein. Therefore, any cluster of interacting residues could be of significant interest from the protein folding and function point of view. The fact that the residues, which form the folding nucleus are also conserved in the sequences, adds to the belief that for each protein structure, there exist a small number of conserved residues that are the key determinants of the folding process and more so these key residues are clustered in the native structure of the protein.[60,61]

Hydrophobic clusters, which form the buried core of the protein, are known to be important in stabilizing the native structure. An alteration of residues in these clusters is known to destabilize the protein.[62] The hydrophobic interactions on the surface of the protein are major determinants of thermal stability. Hydrophobic clusters on the protein surfaces are known to be important for protein oligomerization and subunit association and also play a crucial role in protein–protein and protein–DNA interactions.[63–65] On the other hand, charged clusters are known to be essential for the function of the protein[66–68] and they are mainly present at the active site and metal binding sites of the protein.

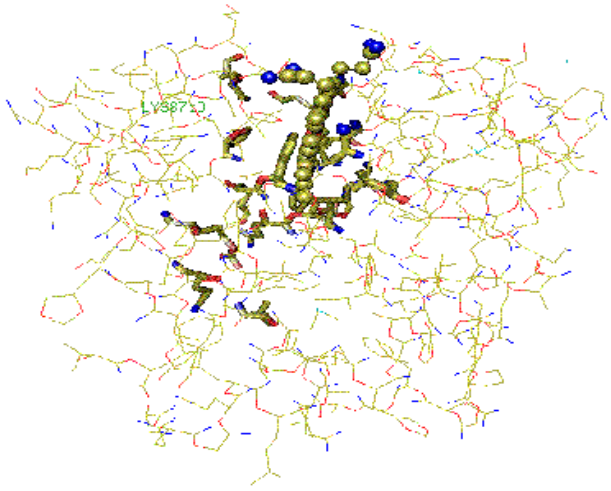
Since residue clusters play a crucial role in protein stabilization, protein–protein association, folding and function, efficient techniques are required to delineate and characterize them from native structures and several methods are available for this purpose.[69–71] Such important clusters are identified by the graph spectral technique described earlier and it has proved to be elegant and efficient in cluster identification. The elegance of graph spectral method is due to the fact that it takes a global view of protein structure and the efficiency is due to a single numeric computation, which is not influenced by the starting procedure (like in some matrix reorganization methods) and can handle weighted matrix, unlike some traditional methods dealing with matrices. Some of the specific applications are discussed below.

5.2.1. *Active site clusters*

It is known that the active site of proteins generally consist of a network of interacting residues. The cluster of these interacting residues can be identified using the clustering algorithm. This algorithm can be used to find side chain clusters in proteins and amongst these, the active site clusters can be identified. Kannan and Vishveshwara have shown that one of the clusters obtained using high contact criterion (means a strong overlap between interacting residues), generally belongs to the active site and on reducing the contact criterion, these active site clusters expand.[22] These active site clusters are found to branch out into bigger networks when the contact criterion is further reduced. The expansion of these active site clusters is significant because such network of interactions surrounding the active site could be important for holding the active site residues in proper orientation and also for movement of these residues during ligand binding. Figure 10 shows the clusters obtained in myoglobin (4mbn) at high and low contact criterion.



(a) Myoglobin (4mbn) at high contact criterion.



(b) Myoglobin (4mbn) at low contact criterion.

Fig. 10. Expansion of the active site cluster in myoglobin (pdb code: 4mbn). The side chain clusters obtained in 4mbn at high and low cut off is shown in a and b respectively. The cluster residues along with the porphyrin ring at the active site are shown in bold.

It is evident that the cluster obtained at high cut off has got expanded at low cut off, thus including many other residues in the cluster during the process of expansion. The biological significance of such an expansion of active site cluster is two fold. Primarily, it explains how functionally important residues are appropriately placed in space by being anchored to the core of the protein. Secondly, the nucleation-condensation hypothesis of protein folding[72] seems to be supported.

5.2.2. *Folding clusters*

During protein folding, the first residue–residue interactions occur amongst residues involved in the nucleation sites. The identification of such residues will give some insight to the folding pathway.[72] The sites of nucleation that help in protein folding have been identified using graph-theoretic method. It is well known that hydrophobic interactions play a major role during the process of protein folding.[60,61] Hence, hydrophobic clusters were determined in a set of proteins for which experimental information regarding residues important for folding was already available. The vector components of the top eigenvalues (which in a way represents the weight of the vertex in the graph) were found to correlate very well with the established experimental results regarding the residues that play a major role during the folding of the protein. Hence, the analysis of the vector components of the top eigenvalues of the hydrophobic clusters could give information regarding the folding of the protein. The probable nucleation site on protein triose phosphate isomerase (TIM barrel) was identified using this method.[73] Conserved hydrophobic clusters with high hydrophobicity index[74,75] were identified as the folding nucleus in this class of proteins.

5.2.3. *Topological characterization of α/β barrel fold*

It is a well-known fact that the α/β barrel is one of the commonest folds in proteins and is adopted by a variety of proteins. Though these proteins have similar fold, they scan highly diverse sequence and functional space. This had motivated many researchers into probing the reason for such a structural similarity in spite of large sequence diversity. In order to understand the structural factors contributing to such a unique fold, the structures of these proteins in terms of the backbone clusters, which might have contributed to the formation of α/β barrel framework, have been analyzed.[73] Graph-theoretic algorithm was used to identify backbone clusters,[23] as backbone interactions seem to be more conserved than side-chain interaction in this class of proteins. An interesting result of this study was the observation that the largest eigenvalue (lev) of graphs constructed for topologically similar proteins (α/β barrel proteins in this case)

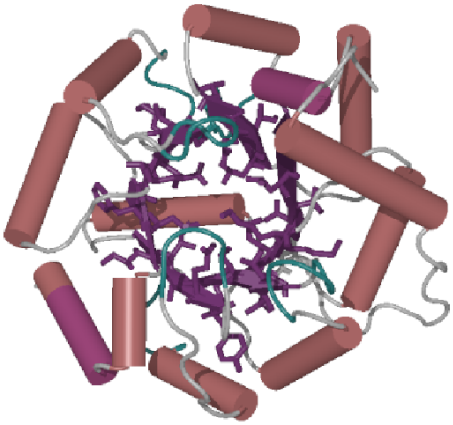
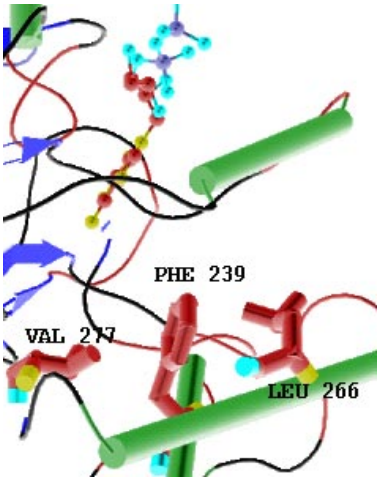


Fig. 11. Backbone clusters in α/β barrel protein, indole-3-glycerolphosphate synthase (1juk). The cluster residues are shown in bond representation. The clusters are concentrated in the β-barrel region and the loops connecting the β-strands to the α-helices.

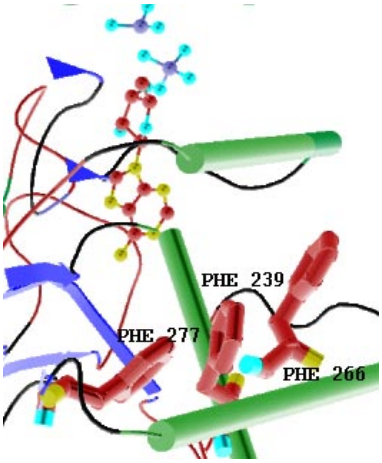
had similar values. The lev depends upon the extent of branching in the graph and seemed to capture the topology of a protein fold. In the case of α/β barrel proteins, the residues contributing to the largest vector components corresponding to the largest eigenvalue [lvc(L)], seem to belong to the β-barrel region, indicating a possible structural role of these residues. More importantly, these lvc(L) residues are topologically conserved in all α/β barrel proteins studied. The residues at the middle and C-terminal ends of the strands contribute significantly to the cluster formation (largest vector component magnitude) and are also located close to the active site of the proteins. Figure 11 shows the backbone clusters obtained in indole-3-glycerolphosphate synthase (1juk), which is an α/β barrel protein. It is evident that the clusters are concentrated in the β barrel region and that the major contributions are from residues in the strands and loops regions. It would be interesting to see if topological indices such as lev or the graph spectra can be used to characterize protein topologies.

5.2.4. *Aromatic clusters/thermal stability*

Thermophilic proteins can exist in their native conformation and be functionally active at high temperatures as well. A number of factors such as hydrogen bonds and salt bridges have been known to be the determinants of thermal stability.[76–78] Cluster analysis on a dataset of thermophilic proteins showed an over-representation of aromatic clusters/interactions in thermophilic proteins as compared to their mesophilic homologue.[79] Moreover, most of these additional aromatic clusters found in thermophiles were located on the surface of the protein and were in relatively rigid regions of the protein showing low B-factors. In the mesophilic counterpart, these additional aromatic



(a) Mesophilic phosphoglycerate kinase (3pgk).



(b) Thermophilic phosphoglycerate kinase (1php).

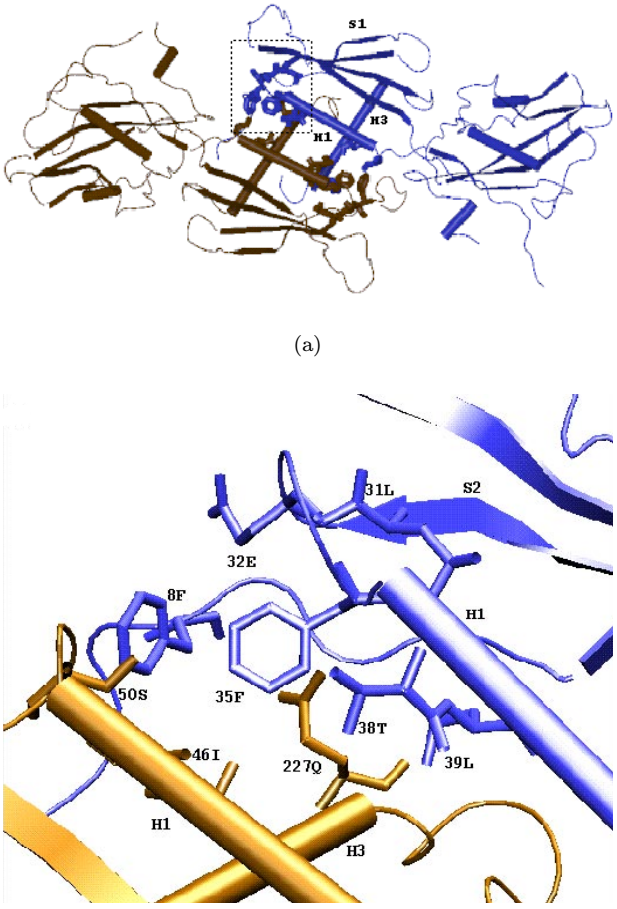
Fig. 12. Aromatic clusters in mesophilic (3pgk) and thermophilic (1php) phosphoglycerate kinase. Residues 239F, 266L and 277V from a cluster in the mesophile. The non-aromatic leucine and valine are replaced by aromatic phenylalanines in positions 266 and 277 respectively, in the thermophilic protein, thus forming a network of aromatic interactions.

residues, are substituted by non-aromatic residues. These features can be seen in the example of the protein phosphoglycerate kinase as shown in Fig. 12. This finding provides a basis for experimentally verifying the role of aromatic interactions in contributing to protein thermal stability.

5.2.5. *Protein–protein interaction*

Multimeric proteins are extremely common in nature. The interactions, which hold the monomers of a multimeric protein are usually non-covalent in nature and form the basis for protein–protein interaction and protein assembly. Analysis of the nature of interactions in such interfaces is critical in understanding the structural factors underlying protein–protein recognition. Since the structural environment of residues involved in protein–protein interfaces would be similar to residues buried in a protein core,[80] one would expect similar stabilizing interactions governing proteins interior and interfaces. In order to verify this hypothesis, a set of homodimeric interfaces was analyzed for the presence of side-chain clusters at the interface.[81] The analysis showed that there are specific amino acid residues, which cluster at the interface, involved in strengthening the monomer–monomer interaction. The residues which form the center of the interface clusters [identified using the lvc(L)] were highly conserved and are more likely to disrupt the dimer interface upon mutation. Thus identifying side chain clusters and their centers (nodes with highest degree usually have the largest vector component and generally represent the center of the cluster in an approximate spherical distribution) using graph theory has helped in identifying "hot spots" (residues important for dimer stability, the mutation of which destabilizes the process of oligomerization) at protein interfaces. The graph-spectral algorithm in conjunction with residue conservation and other traditional methods like determination of the loss in accessible surface area on dimerization, was used to predict potential dimerization sites in monomeric proteins. The predicted "hot spots" and dimerization sites correlate well with the experimental results and can serve as a powerful tool for protein structure analysis and creating protein interaction networks.

Further, the procedure was used as a predictive tool in the case of RNA polymerase. The α_2 dimer of core RNA polymerase forms the initiation step in the assembly of complete protein. Two "hot spots" at the N-terminal domain of the α-subunit were predicted to be important for dimer stabilization by the graph spectral algorithm. Specifically $F35$ formed the center of the cluster and was predicted to be critical in dimerization. This prediction was experimentally verified using site-directed mutagenesis[82] and was shown to be dimerization defective. Figure 13 shows the



(a)

(b)

Fig. 13. Side chain clusters at the interface of RNA polymerase α_2 dimer. The interface cluster enclosed within the rectangular box in (a) has been enlarged in (b). The cluster forming residues at the interface include 46I, 50S, 8F, 32E, 35F, 31L, 38T and 39L from one monomer and 227Q from the other monomer. Residue 35F has been identified as the "hot spot" using the graph spectral method and has been experimentally verified to be important for dimerization.

cluster identified in the dimer interface of RNA polymerase α_2 dimer. The cluster consists of residues from both monomers and hence, stabilizes the interface through a network of interactions.

5.3. *Graph theory in protein structure and folding*

5.3.1. *Comparative modeling*

An exponential increase in genome sequencing has resulted in the availability of a large number of protein sequences. Ultimately, it is of interest to know how these proteins function, for which the three-dimensional knowledge of structure plays an important role. With an increase in the number of protein crystal structures, it has been possible to catalog most of the possible folds taken by proteins in nature. Rigorous developments have taken place in assigning a given sequence to a known fold.[83] The fold reveals the overall topology taken up by the polypeptide chain. However, the possible combinations of side chain conformations and the non-regular secondary structures of the backbone conformations are astronomical. Conformational comparison with homologous proteins decrease the search space (known as comparative modeling). Clique finding algorithm is made use of by Samudrala and Moult,[84,85] to explore the side chain conformational space in comparative modeling. Weighted nodes and edges are considered for graph construction in these studies. The weight of each node is based on the degree of interaction between the main chain and the side chain. The nodes are possible conformations of the side chains. Edges are drawn between all considered nodes with the exception of nodes arising from the same side chain and nodes giving rise to steric clashes. The edges are also weighted based on the extent of favorable interaction between nodes. Once the entire graph is constructed, all the maximal sets of completely connected nodes are detected by clique finding algorithm.[86] The cliques are scored based on the node and edge weights. The algorithm is used in comparative modeling to build side-chains, segments of main-chain and mix and match between different homologues in a context sensitive manner.

5.3.2. *Inverse folding*

The problem of inverse folding is to identify all possible sequences, which can take up a given structure.[87] In other words, it amounts to the design of sequences for a given structure. Lattice models are simple to handle and it allows one to completely enumerate the energetics of all possible conformations taken up by a sequence and vice versa. The possible application of graph theoretic parameters to design sequences for a chosen structure has been explored.[88] A compact conformation on a lattice was chosen with the polymer chain on lattice points as vertices and non-sequential neighbors as edges. Hypothetical sequences with polar (P) and hydrophobic (H) residues were chosen as the polymer chain and different energy values were assigned for (H–H), (H–P) and (P–P) non-sequential contacts. Adjacency matrix of the graph obtained for a given conformation was constructed and the eigenvalues and their corresponding vector components were evaluated. It was seen that the vector components of the largest eigenvalue were able to discriminate between vertices of same degree depending on its position in the graph. This feature was exploited to assign topological weights for the vertices, which in turn assisted in placing a H or a P residue on the chosen vertex for the chosen conformation to take up lowest energy. The procedure worked quite well to construct sequences for a chosen conformation on lattice model. The usefulness of such topological indices in designing realistic sequences for real protein structures is yet to be explored.

5.3.3. *Searching for the rules of protein folding*

Whether the structures of proteins have evolved randomly or whether there are hidden constraints on their patterns, scope and complexity, is a puzzle, which is being investigated[89] in the field of protein folding. Przytycka *et al.*, have studied protein sequences and structures to see if patterns in protein structures are haphazard assortments or whether they are similar to sentences in a language which can be generated by an underlying grammar.[42] They have considered the motifs of β-proteins like β-hairpin, β-β-β unit, the anti-parallel and parallel Greek key motifs (some of which are shown in Fig. 6) as building blocks, and a set of simple rules (hypothetical) such as chirality,

non-crossing properties and preservation of hydrogen bond patterns in β-sheet topologies, to generate all known β-folds in nature. The studies are carried out on β-proteins and can be extended to other class of proteins as well. The β-strand regions are identified from the secondary structural assignment,[49] and they are considered as nodes of the graph. To begin with edges are formed between sequentially adjacent nodes and the Adjacency matrix is recursively modified by applying the simple rules of folding mentioned above. The process is continued until the complete domain and its connections are identified. By this procedure, it was possible to generate the protein β-folds found in nature, which led them to conclude that there indeed is an underlying folding grammar in the evolution of protein structures. The complete and correct set of rules however has to be established by further investigations.

5.4. *Protein dynamics*

A protein should not only fold to an unique conformation, but also should posses the ability to do conformational motions relevant to its function, while retaining its folded state. It is likely that the molecular topology inherently determines the most likely mechanism of motion in a cooperative manner involving all parts of the structure. The rigorous method of understanding protein dynamics is to carry out molecular dynamics (MD) simulations (or Monte Carlo simulations) where the force on each atom is explicitly defined and the subsequent position of the constrained atoms are determined by Newton's law of motion. This procedure however is computationally intense since the simulations have to be carried out for several nanoseconds to obtain the equilibrium properties. Further, the procedure becomes prohibitive for large structures. A molecule with N atoms in principle can have $3N$ degrees of freedom. However, the atoms in a molecule are constrained by various types of forces, such as covalent, hydrogen bond, hydrophobic and so on. As a result, the essential degrees of freedom are reduced to much less than $3N$. Essential dynamics studies which identify the important modes in proteins is similar to normal mode analysis of small molecules. The method such as Gaussian Network Model (GNM)[35] and constrained graph theory,[36] which are described below, have exploited the feature that information regarding protein dynamics is encoded in the molecular topology and have developed methods to extract such information.

5.4.1. *GNM model*

The Gaussian network model (GNM), developed by Bahar and coworkers,[35,90] yields dynamical characteristics of biomolecular structures based on atomic coordinates of the native conformation. The basic assumption of the model is to represent the native state as a perfect elastic network. The amino acid residues (represented by C-α atoms) are the nodes and the spatially close residues are connected as edges, which represent the spring-like interactions between the nodes. Edges are defined between all those residues (including the immediate sequence neighbors) which reside within a sphere of 7Å as shown in Fig. 7. The Laplacian (Kirchoff) matrix is constructed based on this definition and the solution to such a matrix is obtained as eigenvalues, a procedure similar to the one described earlier. The graph spectra thus obtained is made use of in extracting interesting dynamical properties from the principles of statistical thermodynamics.

The connectivities are considered as springs and the total potential of the molecular system of N sites is given as the sum of pairwise interactions

$$V_{\text{tot}} = \sum_i \sum_j V(R_i R_j) = \sum_i \sum_j \left(\frac{1}{2}\right) \gamma(\Delta R_{ij} \cdot \Delta R_{ij}) \qquad (12)$$

between all pairs $1 \leq i < j \leq N$. Here, γ is the Hookean force constant, $\Delta R_{ij} \equiv \int R_{ii} - R^0_{ij}$ is the fluctuation in the distance vector $R_{ij} = R_j - R_i$ between sites i and j, relative to the equilibrium separation R^0_{ij}. Using the concepts of statistical mechanics, expressions are derived for mean square fluctuation, based on the trace of the Laplacian matrix. The temperature factor (or B factor) which can be experimentally obtained from X-ray crystallography is related to this mean square fluctuation. Such calculations are carried out on a series of proteins and their complexes have been found to predict values consistent with experiment. Similarly, vibrational contributions to the Helmholtz free energy of the protein is derived based on the eigenvalues of the Laplacian matrix.[91]

The free energy cost of distorting residues on a local scale is thus computed and is applied[35] to obtain information on Proton/Deuterium exchange under weakly denaturing conditions, a parameter which can be determined by $2d$-NMR experiments. Thus, experimentally observed dynamics results are correlated with the results from the superposition of N-1 modes in the GNM representation of proteins, the frequencies of which are given by non-zero eigenvalues, $\lambda_k (2 \leq k \leq N)$, of the Laplacian matrix. The results from the calculations have been shown to agree with the hydrogen exchange data on a number of proteins such as cytochrome C, ribonuclease H and so on.

The slowest vibrational mode usually extends over the entire molecule and thus represents the global or most cooperative mode. The distribution and the frequency of a given mode is given by its eigenvalue and its vector component respectively. Analysis of these parameters of low modes gives valuable dynamical information on the global motions of amino acid residues in proteins. For instance, the mode shapes are calculated for the antibody, immunoglobulin $G(IgG)$.[35] The hinge regions are shown to be rigid, whereas the variable domains, which recognize the antigen, are computed to be flexible, which is consistent with its biological activity. The correlation of global motion with biological activity is also tested for other proteins such as HIV reverse transcriptase. Recently, the molecular mechanism of GroEL-GroES chaperonin complex, which prevents misfolding of proteins, has been investigated using this method.[92] The relative flexibility of the hinge region, stability of central cavity and co-operative cross-correlation between subunits of such a large protein–protein complex have been investigated.

5.4.2. *Protein flexibility by constrained graph theory*

The rigidity theory in mathematics was traditionally applied to solve problems in engineering such as structural stability of different truss configurations in bridges. Constrained graph theory was developed for analyzing the rigidity of substructures within covalent network glasses,[93] which was later applied to protein structures.[36] The fundamental step on which the calculations are based is the ability to test whether a constraint is redundant or independent. Such algorithms help in identifying the rigid and flexible substructures in proteins taking into account constraints such as covalency and hydrogen bonds. Protein network is constructed with edge weights related to the force constant between the atoms. The eigenvalues of such a matrix was used to calculate the normal mode frequencies of the network, from which the flexible and rigid substructures are identified. A flexibility index was introduced as a continuous measure for quantifying the flexibility or stability of each bond within the protein. The straightforward evaluation is computationally intense and efficient algorithms have been implemented for fast computing. The methodology is applied to analyze several interesting proteins such as HIV protease and dihydrofolate reductase and the changes in the flexible regions upon drug binding have been detected. The advantage of this method and the GNM model is that they capture the functionally important conformational flexibility when performed on a single structure.

6. Future directions

For structural biologists, there is an explosion of protein structure data and exponential increase in computing power. Although the principles of graph theory were known more than a century ago, structural biologists are now finding exciting applications of this branch of mathematics and there is a promise of this technique to contribute substantially more to our understanding of the problem of protein structure, folding, stability, function and dynamics.

Some of the problems that can be explored include the development of topological indices based on the spectra of protein graphs so as to identify and classify protein folds in the database. The utility of indices can also be explored in protein structure prediction and protein folding simulations. Further, investigations can also be taken up to deduce clustering information from protein sequences based on the conservation of the cluster residues. The graph theoretical approach which captures the global connectivity in a protein molecule can be used to identify important side-chain networks which are involved in transducing energy in a protein molecule.[94] Identifying such

structural networks can help in understanding how these molecular machines work. Research in some of these areas is in progress in our laboratory.

Appendix A
Clustering from Laplacian Matrix[34]

Given n points and $n \times n$ symmetric Adjacency matrix A_{ij} which gives the connection between points i and j, we want to find the location of "n" points which minimizes the weighted sum of the squared distances between the points.

If x_i denotes the X coordinate of point "i" and Z denotes the weighted sum of the squared distances between the points,

$$Z = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 A_{ij} \tag{13}$$

where A_{ij} is the Adjacency matrix, then the one-dimensional problem is to find a row vector $X' = (x_1, x_2, \ldots, x_n)$ which minimizes the above function where prime denotes the vector transposition. To avoid the trivial solution $x_i = 0$ for all i, the following quadratic constraint is imposed.

$$X'X = 1\,. \tag{14}$$

The solution to the above framed problem is as follows.

Now it can be shown that the Eq. (13) can be rewritten in terms of the Laplacian matrix as

$$Z = X'LX\,. \tag{15}$$

Expanding Eq. (13), we get

$$Z = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i^2 - 2x_i x_j + x_j^2)A_{ij} \tag{16}$$

$$Z = \sum_{i=1}^{n} x_i^2 a_i - \sum_{j=1}^{n}\sum_{i\neq j}^{n} x_i x_j A_{ij}\,. \tag{17}$$

Since A_{ij} is a symmetric matrix $a_i = a'_j$, Z can be written as in Eq. (15).

To minimize Z subject to the constraint $X'X = 1$, introduce the Lagrangian multiplier λ and form the Lagrangian

$$L = X'LX - \lambda(X'X - 1)\,. \tag{18}$$

Taking the first partial derivative of L with respect to the vector X and setting the result equal to zero yields

$$2LX - 2\lambda X = 0\,. \tag{19}$$

If I is identified as the identity matrix, Eq. 19 can be rewritten as:

$$(L - \lambda I)X = 0 \tag{20}$$

which yields a nontrivial solution X, if and only if λ is an eigenvalue of the matrix L and X is the corresponding eigenvector. If the above equation is premultiplied by X' and the constraint Eq. (14) is applied, we obtain

$$\lambda = X'LX\,. \tag{21}$$

Thus, the formal solution to Eqs. (14) and (15) is simply that X is the eigenvector of L, which minimizes Z and λ is the corresponding eigenvalue. The minimum eigenvalue zero yields the uninteresting solution $X = (1/\sqrt{n}, 1/\sqrt{n}, \ldots, 1/\sqrt{n})$. Hence the second smallest eigenvalue and the associated eigenvector which yields the optimal solution is considered. We can that this solution is related to the clustering of points. The above solution for one dimension also holds good in two- and three-dimensional space.[34]

Appendix B
Definition of nodes and edges for clustering algorithm[22]

The protein graphs are constructed using $C\beta$ atoms of the amino acids as nodes and the distance between the $C\beta$ atoms as edges if the specified interaction criterion is satisfied. The side chain interaction criterion between two amino acid residues is evaluated by using an expression similar to that used by Heringa and Argos.[69] The expression is of the following form:

$$INT(R_i, R_j) = \frac{N(R_i, R_j)}{\mathrm{Norm}(\mathrm{Restype}(R_i)) \times 100} \tag{22}$$

where $N(R_i, R_j)$ is the number of distinct interacting pairs of side chains atoms between the residues R_i and R_j. If any two side chain atoms of residues R_i and R_j are within a distance of 4.5 Å, then they are said to form an interacting pair. All such interacting pairs between residues R_i and R_j are counted to obtain $N(R_i, R_j)$.

The normalization values (Norm(Restype(R_i))) for all 20 residue types R_i was obtained by the following expression,

$$\text{Norm}(\text{Restype}(R_i)) = \frac{\sum_{K=1}^{p} \text{Maxm}(\text{Type}(R_{ik}))}{p}. \tag{23}$$

In order to evaluate the normalization factors, an analysis on the non-redundant data set[95] of 148 proteins with a resolution greater than 2.0 Å was performed. The number of interaction pairs (both main chain and side chain) made by the residue type R_i with all its surrounding residues in a protein k was evaluated. Maxm(Type(R_{ik})) was determined by the maximum number of interactions made by residue R_i in protein k. For example, if residue type alanine occurred twice in protein k and if one alanine had 10 interaction pairs with the main chain and side chain atoms of the surrounding residues and the other alanine 12 interaction pairs, then Maxm(ALA_k) is equal to 12. In the same manner, Maxm(Type(R_{ik})) for residue R_i was evaluated for each of the proteins k in the dataset. Norm[Restype(R_i)] was obtained by the average of the maximum interaction value of the residue R_i, over all the proteins p in the dataset, in which the residue type R_i had occurred. The same procedure was followed to obtain the normalization values for all the 20 residue types. The normalization values obtained are given in Table 5. It can be observed that these values correlate well with the size of the amino acid residue.

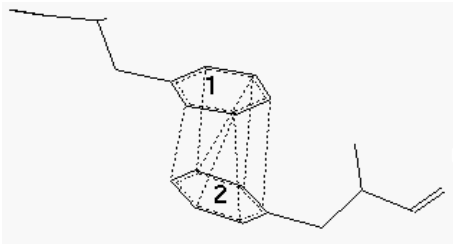
When interaction between all side chains (polar and non-polar) are considered for constructing the graph, the overlap criterion can be classified as follows:

(a) If the percentage interaction between the two side chains is 8% or more, then it is defined as high side chain overlap.
(b) If the percentage interaction between the two side chains is more than 5% and less than 8%, then it is defined as medium side chain overlap.
(c) If the percentage interaction is less than 5%, then it is defined as low side chain overlap.

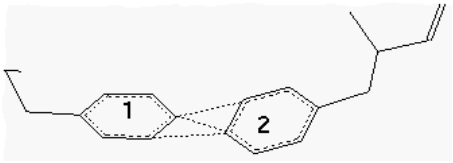
This has been elucidated in Fig. 14.

Table 5. Normalization value used in the evaluation of the percentage contact between pairs of interacting amino acids as given in Appendix B. (The values for the 20 amino acid residue type derived from 148 protein structures).

S. No	Residue Type	Norm
1	Alanine	55.7551
2	Arginine	93.7891
3	Asparagine	73.4097
4	Aspartic acid	75.1507
5	Cystine	54.9528
6	Glutamine	78.1301
7	Glutamic acid	18.8288
8	Glycine	47.3129
9	Histidine	83.7357
10	Isoleucine	67.9452
11	Leucine	72.2517
12	Lysine	69.6096
13	Methionine	69.2569
14	Phenyl alanine	93.3082
15	Proline	51.3310
16	Serine	61.3946
17	Threonine	63.7075
18	Trptophan	106.703
19	Tyrosine	100.719
20	Valine	62.3673



(a) High contact (8.57%)



(b) Low contact (3.21%)

Fig. 14. High and low contact criteria. Two pairs of phenylalanine rings interacting with each other are shown. The dotted lines between the phenylalanines indicate the atoms that are within a distance of 4.5 Å. (a) high contact (8.57%); (b) low contact (3.21%). The method of evaluating the percentage contact is described in Appendix B.

Acknowledgments

We wish to thank B.S. Sanjeev and S.M. Patra for their contributions to the development of this subject in our lab. One of the authors (KVB) would like to thank the Centre for Scientific and Industrial Research (CSIR), India, for the fellowship offered.

References

1. N. Deo, *Graph Theory with Applications to Engineering and Computer Science*, 2nd edn., F. George (Prentice Hall of India Private Limited, New Delhi, 1984).
2. O. Ivanciuk and A.T. Balaban, *Encyclopaedia of Computational Chemistry, Graph Theory in Chemistry*, Vol. 2, ed. P.V.R. Schleyer (John Wiley & Sons, New York, Singapore, 1998).
3. I. Gutman and O.E. Polansky, *Mathematical Concepts in Organic Chemistry* (Springer-Verlag, Berlin, 1986).
4. D. Bonchev and D.H. Rouvray, *Chemical Graph Theory: Theory and Fundamentals* (Abacus/Gorden & Breach Science, New York, 1991).
5. N. Trinajstic, *Chemical Graph Theory*, 2nd edn. (CRC Press, Boca Raton, FL., 1992).
6. S.J. Cyvin, *Tetrahedron Letters* **22**(28), 2709 (1981).
7. G.W.E. Milne, *J. Chem. Inf. Comput. Scie.* **41**(3) (2001).
8. D. Bonchev, A.T. Balaban and O. Mekenyan, *J. Chem. Inf. Comput. Sci.* **20**, 106 (1980).
9. N. Biggs, *Algebraic Graph Theory* (Cambridge University Press, Cambridge, 1974).
10. G.V. Strang, *Linear Algebra and its Applications* (Harcourt Brace Jovanovich, San Diego, 1988).
11. P. Gould, *Inst. Br. Geog. Trans.* **42**, 53 (1967).
12. K. Tinkler, *Inst. Br. Geog. Trans.* **55**, 17 (1972).
13. D. Cvetkovic, M. Doob, I. Gutman and A. Torgasev, *Recent results in the Theory of Graph Spectra, Annals of Discrete Mathematics*, ed. P.L. Hammer (Elsevier Science Publishers B.V., Amsterdam, Netherlands, 1988).
14. D. Cvetkovic and I. Gutman, *Croat. Chem. Acta* **49**, 105 (1977).
15. H. Weiner, *J. Am. Chem. Soc.* **69**, 17 (1947).
16. H. Weiner, *J. Phys. Chem.* **52**, 1082 (1948).
17. H. Hosoya, *Bull. Chem. Soc. Jpn.* **44**, 2332 (1971).
18. A.T. Balaban, I. Motoc, D. Bonchev and O. Mekenyan, *Topics in Current Chemistry* **114**, 21 (1983).
19. M. Randic and S.C. Basak, *J. Chem. Inf. Comput. Sci.* **41**, 650 (2001).
20. B.E. Eichenger and J.E. Martin, *J. Chem. Phys.* **69**(10), 4595 (1978).
21. B.E. Eichenger, *Macromolecules*, **13**(1), 1 (1980).
22. N. Kannan and S. Vishveshwara, *J. Mol. Biol.* **292**, 441 (1999).
23. S.M. Patra and S. Vishveshwara, *Biophys. Chem.* **84**, 13 (2000).
24. J.R. Ullman, *J. Assoc. Comput. Mach.* **23**, 31 (1976).
25. E.M. Mitchell, P.J. Artymiuk, D.W. Rice and P. Willet, *J. Mol. Biol.* **212**(1), 151 (1989).
26. H.M. Grindley, P.J. Artymiuk, D.W. Rice and P. Willet, *J. Mol. Biol.* **229**, 707 (1993).
27. P.J. Artymiuk, A.R. Poirette, H.M. Grindley, D.W. Rice and P. Willet, *J. Mol. Biol.* **243**, 327 (1994).
28. W.H. Haemers, *Linear Algebra Appl.* **3**, 593 (1995).
29. L. Hagen and A.B. Kahng, *IEEE Transactions Comput. Aided Des.* **11**(9), 1074 (1992).
30. E.R. Barnes, *SIAM J. Algebra Discrete Meth.* **7**(2), 541 (1982).
31. R.B. Boppanna, *Proc. IEEE Symp. Found. Comput. Sci. 25th Ann. Symp.*, 280 (1987).
32. W.E. Donath and A.J. Hoffman, *IBM J. Res. Dev.* 420 (1973).
33. http://www.boost.org
34. K.M. Hall, *Manag. Sci.* **17**, 219 (1970).
35. I. Bahar, *Rev. Chem. Eng.* **15**(4), 319 (1999).
36. D.J. Jacobs, A.J. Rader, L.A. Kuhn and M.F. Thorpe, *Proteins: Struct. Funct. Genet.* **44**, 150 (2001).
37. M. Levitt and C. Chothia, *Nature* **261**(5561), 552 (1976).
38. M.J. Sternberg and J.M. Thornton, *J. Mol. Biol.* **110**(2), 269 (1977).
39. J.S. Richardson, *Adv. Protein Chem.* **34**, 167 (1981).
40. I. Koch, F. Kaden and J. Selbig, *Proteins* **12**(4), 314 (1992).
41. I.V. Grigoriev, A.A. Mironov and A.B. Rakhmaninova, *J. Biomol. Struct. Dynam.* **12**, 559 (1994).
42. T. Przytycka T., Srinivasan R. and R.D. Rose, *Protein Sci.* **11**, 409 (2002).
43. S. Miyazawa and R.L. Jernigan, *Macromolecules* **18**, 534 (1985).
44. S. Miyazawa and R.L. Jernigan, *J. Mol. Biol.* **256**, 623 (1996).
45. C. Chothia, *Nature* **357**(6379), 543 (1992).
46. L.L. Conte, B. Ailey, T.J. Hubbard, S.E. Brenner, A.G. Murzin and C. Chothia, *Nucleic Acids Res.* **28**(1), 257 (2000).
47. D. Fischer and D. Eisenberg, *Curr. Opin. Struct. Biol.* **9**(2), 208 (1999).
48. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliand, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne, *Nucleic Acids Res.* **28**, 235 (2000).
49. W. Kabsch and C. Sander, *Biopolymers* **22**, 2577 (1983).
50. N. Srinivasan, R. Sowdhamini, C. Ramakrishnan and P. Balaram, *Molecular Conformation and Biological Interactions*, eds. P. Balaram and S. Ramaseshan (Indian Academy of Sciences, Bangalore).
51. G. Levi, *Calcolo* **9**, 341 (1972).

52. J. J. McGregor, *Software Pract. Exp.* **12**, 23 (1982).
53. C.W. Crandell and D.H. Smith, *J. Chem. Inf. Comput. Sci.* **23**, 186 (1983).
54. D. Goldman, S. Istrail and C. Papadimitriou, *Extended Abst. Proc. 40th FOCS*, 512 (1999).
55. A.K. Ghose and G.M. Crippen, *J. Comput. Chem.* **6**, 350 (1985).
56. A.T. Brint and P. Willet, *J. Chem. Inf. Comput. Sci.* **27**, 152 (1987).
57. P.J. Artymiuk, H.M. Grindley, J.E. Park, D.W. Rice and P. Willet, *FEBS Lett.* **303**, 48 (1992).
58. P.J. Artymiuk, H.M. Grindley, K. Kiran, D.W. Rice and P. Willet, *FEBS Lett.* **324**(1), 15 (1993).
59. P.J. Artymiuk, D.W. Rice, E.M. Mitchell and P. Willet, *Protein Eng.* **4**, 39 (1990).
60. E. Shanknovich, V. Abkevich and O. Ptitsyn, *Nature* **379**, 96 (1996).
61. L. Mirny, V. Abkevich and E. Shanknovich, *Proc. Natl. Acad. Sci. USA* **283**, 507 (1998).
62. D. Pins, D.E. Anderson, W.A. Baase, F.W. Dahlquist and B.W. Matthews, *Biochemistry* **30**, 11521 (1991).
63. C. Chothia and J. Janin, *Nature* **256**, 705 (1975).
64. L. Young, R.L. Jernigan and D.G. Covell, *Protein Sci.* **3**(5), 717 (1994).
65. J.E. Anderson, M. Ptashne and S.C. Harrison, *Nature* **326**, 846 (1987).
66. S. Karlin and Z.Y. Zhu, *Proc. Natl. Acad. Sci. USA* **93**, 8344 (1996).
67. S. Kanaya, C. Katsuda-Nakai and M.J. Ikehara, *J. Biol. Chem.* **266**(18), 11621 (1991).
68. M.D. Davis, R.S. Wonderling, S.L. Walker and R.A. Owens, *J. Virol.* **73**(3), 2084 (1999).
69. J. Heringa and P. Argos, *J. Mol. Biol.* **220**, 151 (1991).
70. M. H. Zehfus, *Protein Sci.* **4**, 1188 (1995).
71. M. B. Swindells, *Protein Sci.* **4**, 93 (1995).
72. A.R. Fersht, *Curr. Opin. Struct. Biol.* **7**, 3 (1997).
73. N. Kannan, S. Selvaraj, M. Gromiha and S. Vishveshwara, *Proteins: Struct. Funct. Genet.* **43**, 103 (2001).
74. Y. Nozaki and D. Tanford, *J. Biol. Chem.* **246**, 2211 (1971).
75. D.D. Jones, *J. Theor. Biol.* **50**, 167 (1975).
76. R. Jaenicke and G. Bohm, *Curr. Opin. Struct. Biol.* **8**, 738 (1998).
77. R. Ladenstein and G. Antranikian, *Adv. Biochem. Eng. Biotechnol.* **61**, 37 (1998).
78. A. Szilagyi and P. Zavodsky, *Structure* **8**, 493 (2000).
79. N. Kannan and S. Vishveshwara, *Protein Eng.* **13**, 753 (2000).
80. S. Jones and J. M. Thornton, *Proc. Natl. Acad. Sci. USA* **93**(1), 13 (1996).
81. K.V. Brinda, N. Kannan and S. Vishveshwara, *Protein Eng.* **15**(4), 265 (2002).
82. N. Kannan, C. Preethi, G. Pallavi, S. Vishveshwara and C. Dipankar, *Protein Sci.*, **10**, 46 (2001).
83. E.V. Koonin, Y.I. Wolf and L. Aravind, *Adv. Protein Chem.* **54**, 245 (2000).
84. R. Samudrala and J. Moult, *J. Mol. Biol.* **279**(1), 287 (1998).
85. R. Samudrala and J. Moult, *Proteins: Struct. Funct. Genet.* **1**, 43 (1997).
86. C. Bron and J. Kerbosch, *Communi. ACM*, **16**(9), 575 (1973).
87. J.W. Ponder and F.M. Richards, *J. Mol. Biol.* **193**(4), 775 (1987).
88. B.S. Sanjeev, S.M. Patra and S. Vishveshwara, *J. Chem. Phys.* **114**(4), 1906 (2001).
89. J.R. Banavar, A. Flammini, D. Marenduzzo, A. Maritan and A. Trovato, *Proteins: Struct. Funct. Genet.*, **47**, 315 (2002).
90. I. Bahar, A.R. Atilgan and B, Erman, *Fold. Des.* **2**, 173 (1997).
91. I. Bahar, A.R. Atilgan, M.C. Demirel and B. Erman, *Phys. Rev. Lett.* **80**, 2733 (1998).
92. O. Keskin, I. Bahar, D. Flatow, D.G. Covell and R.L. Jernigan, *Biochemistry*, **41**(2), 491 (2002).
93. M.F. Thorpe, D. Jacobs, N. Chubynsky and A. Rader, "Generic Rigidity of Network Glasses," in *Rigidity Theory and Applications*, eds. M.F. Thorpe and P. Duxbury, (Kluwer Academic/Plenum, New York, 1999).
94. S.W. Lockless and R. Ranganathan, *Science* **286**, (5438), 295 (1999).
95. U. Hobohm and C. Sander, *Protein Sci.* **3**, 522 (1994).
96. W. Humphray, A. Jalke and K. Schutter, *J. Molec. Graphics* **14**(1), 33 (1996).