Article No. jmbi.1999.3058 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (1999) 292, 441–464

JMB

Identification of Side-chain Clusters in Protein Structures by a Graph Spectral Method

N. Kannan and S. Vishveshwara*

Molecular Biophysics Unit
Indian Institute of Science
Bangalore, 560 012, India

This paper presents a novel method to detect side-chain clusters in protein three-dimensional structures using a graph spectral approach. Protein side-chain interactions are represented by a labeled graph in which the nodes of the graph represent the C^{β} atoms and the edges represent the distance between the C^{β} atoms. The distance information and the non-bonded connectivity of the residues are represented in the form of a matrix called the Laplacian matrix. The constructed matrix is diagonalized and clustering information is obtained from the vector components associated with the second lowest eigenvalue and cluster centers are obtained from the vector components associated with the top eigenvalues. The method uses global information for clustering and a single numeric computation is required to detect clusters of interest. The approach has been adopted here to detect a variety of side-chain clusters and identify the residue which makes the largest number of interactions among the residues forming the cluster (cluster centers). Detecting such clusters and cluster centers are important from a protein structure and folding point of view. The crucial residues which are important in the folding pathway as determined by Φ_F values (which is a measure of the effect of a mutation on the stability of the transition state of folding) as obtained from protein engineering methods, can be identified from the vector components corresponding to the top eigenvalues. Expanded clusters are detected near the active and binding site of the protein, supporting the nucleation condensation hypothesis for folding. The method is also shown to detect domains in protein structures and conserved side-chain clusters in topologically similar proteins.

Keywords: clusters; hydrophobic; graph theory; Laplacian matrix-eigenvalue protein folding

*Corresponding author

Introduction

Non-bonded side-chain interactions are important for the stability, function and folding of proteins. The role of non-covalent side-chain interactions in stabilizing the mutual orientation of secondary structures has been studied extensively (Chou *et al.*, 1990; Nemethy & Scheraga, 1979; Creighton & Chothia, 1989). Clusters of hydrophobic side-chains on the surface are known to be important for protein-protein recognition (Young *et al.*, 1994; Guss & Freeman, 1983; van de Kamp *et al.*, 1990; Pelletier & Kraut, 1992; Chen *et al.*, 1994), protein oligomerization (Jones *et al.*, 1985; Ponder & Richards, 1987; Mossing & Sauer, 1990) and protein DNA interactions (Anderson *et al.*, 1987). Often a network of charged side-chains is found near the metal binding site and active site (Wright *et al.*, 1969; Weis & Drickamer, 1994; Ng *et al.*, 1996).

Understanding how individual side-chain interactions in a protein molecule cooperate during the process of folding can give us a possible solution to the Levinthal paradox (Levinthal, 1969). However, monitoring side-chain interactions during the folding process is experimentally difficult. Nevertheless, recent protein engineering methods have been successful in probing the contribution of individual side-chain interactions, to the stability of folding intermediates and transition states (Fersht, 1997). Also NMR techniques have been used to observe the clustering of hydrophobic side-chains during the early stages of folding (Lumb & Kim, 1994).

E-mail address of the corresponding author:
sv@mbu.iisc.ernet.in

0022-2836/99/370441–24 $30.00/0

Identifying specific side-chain clusters that might be formed during the early stages of folding from the analysis of native structure is of considerable importance (Engelhard & Evans, 1996). There are a few methods for such cluster detection reported in the literature. For instance, the method of Heringa & Argos (1991) detects side-chain clusters based on the extent of side-chain interactions. The interaction between the side-chains is evaluated by the number of side-chain atoms which come in close proximity. A constraint on residues forming the cluster to have higher cumulative contacts within themselves than with the rest of the protein results in highly compact clusters of small size. A method for detecting such compact hydrophobic clusters has been proposed (Zehfus, 1995) wherein the most compact set of interacting side-chains are said to form a cluster. Compactness here is measured by dividing the side-chain's solvent accessible surface area by its minimum possible surface area. The detected compact clusters are correlated with protein folding units. Swindells (1995) has proposed a method for detecting hydrophobic cores in protein structures. The method considers only buried hydrophobic residues and detects hydrophobic cores based on non-polar side-chain interactions emanating from different secondary structural elements.

The above mentioned methods are specific to detecting clusters which are compact and hydrophobic and are important from the folding point of view, however charged clusters near metal binding sites and active sites which are functionally important are not necessarily compact or buried. A technique has been described to detect such charged clusters (Karlin & Zhu, 1996) in which residues proximal to the metal ions are mapped to a one dimensional array. Statistically significant amino acid clusters are deduced from the generated linear array.

A mathematically elegant approach which employs graph theoretical techniques has been used here to identify side-chain clusters in protein structures. In the literature, different aspects related to protein structure and sequence have been explored using graph theoretical techniques. For instance, specific side-chain patterns in functionally different proteins have been detected by this approach (Artymiuk *et al*., 1990, 1992, 1994). Techniques derived from graph theory have also been used in comparing secondary structural motifs (Mitchell *et al*., 1990) and analysis of sheet topologies (Koch *et al*., 1992). Recently, algorithms have been presented for protein structure prediction (Samudrala & Moult, 1997) and protein modeling (Samudrala & Moult, 1998). Clusters in 2D and 3D lattice models using graph theory have been investigated (Patra & Vishveshwara, 1998).

A variety of side-chain clusters in proteins have been detected here using techniques derived from graph spectral theory, a sub-field of graph theory. Graph spectral theory has been used in clustering of circuit net-lists (Hagen & Khang, 1992; Boppana, 1987; Garbers *et al*., 1990), wherein a circuit is represented by a weighted graph. A Laplacian matrix for the weighted graph is constructed (see Appendix) and clustering information is derived from the vector components of the second lowest eigenvalue. In a protein structure, the side-chain interactions are represented by a weighted graph (as described in Algorithm) and the constructed graph is represented by a Laplacian matrix. Clusters are obtained directly from the eigenvector associated with the second lowest eigenvalue of the Laplacian matrix and the side-chains which make the largest number of interactions in a cluster (cluster centers) are obtained from the eigenvectors associated with the top eigenvalues. The spectral method uses global information for identifying clusters. The method is computationally efficient and robust, as only a single numerical computation is required to detect clusters in a given structure. The method is implemented in the form of a program and the output of the program is a two-dimensional plot called the ''Cluster Plot''. The program is user friendly and has options to detect clusters of interest.

The method has been applied to a set of proteins which are well studied from the structure and folding point of view. The detected clusters are mostly buried and hydrophobic. Often a buried charged residue is detected along with the hydrophobic clusters. The identified clusters should be important from a structure and stability point of view as they are formed by interactions emanating from different secondary structures of the protein. A good correlation is observed between the detected clusters and experimentally observed folding intermediates as determined by hydrogen exchange experiments. Clusters near the active and binding site of the protein are detected. When the side-chain interaction criteria is lowered, some of the clusters are found to expand and invariably the cluster close to the active/binding site is one among them. The implications of the expanded clusters on the structure and folding of proteins are addressed. Interesting correlations are observed with the vector components corresponding to the top eigenvalues and the Φ_F values that are obtained by protein engineering methods (Fersht, 1997) to probe the formation of specific side-chain interactions during the transition state of folding. The method is also shown to be useful in detecting protein cores and identifying protein domains.

Algorithm

Clustering by graph spectral methods

A set of points in space can be represented in the form of a graph wherein the points represent the vertices of the graph and the distance between the points represents the edges. The constructed graph can be represented mathematically in the form of a matrix called the Laplacian matrix as described in the Appendix. The diagonalization of such a

matrix yields the eigenvalues and eigenvectors, which are shown to contain information regarding the clustering of points and branching of the points in space. Specifically, the vector components of the second lowest eigenvalue carry the clustering information, i.e. all vector components which belong to a cluster have the same value (Hall, 1970) and the vector components of the top eigenvalues contain the information regarding the branching of the points forming the cluster (Randic, 1975) and cluster centers (Cvetkovic & Gutman, 1977; Patra & Vishveshwara, 1998). This general methodology which has been used in other disciplines such as electrical engineering for obtaining clusters in circuit net-lists (Hagen & Kahng, 1992) has been adopted here for the identification and characterization of clusters in protein structures.

In the next four subsections we describe how a graph and the corresponding Laplacian matrix was constructed for a protein structure followed by an illustration of the method by considering an example of lysozyme.

Constructing a graph and the Laplacian matrix for a protein structure

A protein structure can be visualized as a network of side-chain interactions and a graph for this interacting network can be constructed by considering only the interacting residues (the criteria for interaction is discussed below). The C^{β} atoms of the interacting residues are considered as vertices and the distance between the C^{β} atoms as edges if the specified interaction criteria is satisfied.

As explained in the Appendix, the Laplacian matrix B can be obtained by determining the adjacency matrix $\mathbf{A}_{ij}$ and the degree matrix $\mathbf{D}_{ij}$. The simplest way to construct an adjacency matrix is to assign a value of 1 or 0 to the matrix elements $\mathbf{A}_{ij}$ depending on whether i and j are connected or not in the graph. Here the adjacency matrix is constructed with the weights assigned as below:

$$\mathbf{A}_{ij} = \frac{1}{d_{ij}}$$

(if side-chains of residues i and j interact above the specified interaction criteria):

$$\mathbf{A}_{ij} = \frac{1}{100} \quad \text{otherwise}$$

where d_{ij} is the distance between the C^{β} atoms of the residues i and j. A distance of 100 is assigned if two side-chains do not satisfy the interaction criteria so that the corresponding weight (1/100) becomes close to zero. The degree matrix is constructed as follows:

$$\mathbf{D}_{ii} = \sum_{j=1}^{n} \mathbf{A}_{ij} \text{ if } i = j$$

$$\mathbf{D}_{ij} = 0 \text{ when } i \neq j$$

where n is the order of the matrix. Hence the Laplacian matrix B is given by

$$\mathbf{B} = \mathbf{D} - \mathbf{A}$$

The decision as to whether the side-chains i and j interact while constructing the adjacency and degree matrix is based on the extent of interaction between them. The side-chain interaction can be evaluated in several ways. For instance, an atlas of protein side-chain interaction on 400 pairs of amino acid side-chains has been comprehensively presented (Singh & Thornton, 1991) wherein the criteria for interaction of two side-chains was defined by calculating the closest inter-atomic distance between the two side-chains. If the observed distance between any two side-chain atoms of the residues were less than the sum of their corresponding Van der Waals radii plus one, then the two side-chains were considered interacting.

Here we evaluate the side-chain interaction between two residues by an expression similar to that used by Heringa & Argos (1991). The expression is of the form:

$$INT(R_i,\ R_j) = \frac{N(R_i,\ R_j)}{NORM(RESTYPE(R_i))} \times 100 \qquad (1)$$

where $N(R_i, R_j)$ in the above expression is the number of distinct interacting pairs of side-chain 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 expression of the form:

$$NORM(RESTYPE(R_i)) = \frac{\sum_{k=1}^{p} MAXM(TYPE(R_{ik}))}{p} \qquad (2)$$

In order to evaluate the normalization factors, an analysis on a non-redundant data set (Hobohm & Sander, 1994) 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 residue type R_i with all its surrounding residues in a protein "k" was evaluated. $MAXM(TYPE(R_{ik}))$ was considered 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 ten interaction pairs with the main-chain and side-chain atoms of the surrounding residues and the other alanine 12 interaction pairs, then $MAXM(ALA_k)$ was determined to be 12. Similarly, $MAXM(TYPE(R_{ik}))$ for residue R_i was evaluated for each of the proteins k in the dataset. $NORM(RESTYPE(R_i))$ in equation (2) was obtained by the

average of the maximum interaction value of the residue R_i, over all the data set of proteins p in which 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 1 and it can be observed that the obtained values correlate with the size of the amino acid residue.

Heringa & Argos (1991) had evaluated the interaction term $INT(R_i, R_j)$ in a slightly different way. They evaluated $N(R_i, R_j)$ by counting the number of side-chain atoms of residue R_j, which come in close proximity to residue R_j. This resulted in $N(R_i, R_j)$ to be different from $N(R_j, R_i)$ whereas, the present way of evaluating the side-chain interaction based on the number of interaction pairs results in $N(R_i, R_j)$ to be equal to $N(R_j, R_i)$. The terms in expression (1) and (2) are further elucidated with an example. The side-chain interaction illustrated in Figure 1(a) shows a schematic representation of two phenylalanine rings interacting with eight side-chain atom pairs with a distance of 4.5 Å. Here the number of interacting pairs $N(R_i, R_j) = 8$ and the normalization value $NORM$(PHE) obtained for phenylalanine (Table 1) is 93.308. Using expression (1), $INT(R_i, R_j)$ is evaluated to be 8.3 %. Figure 1(b) is a schematic representation of two phenylalanine rings having a side-chain interaction of 3.2 %.

Clusters of amino acids are known to occur near the active site, in the interface regions of the interaction between protein-protein and protein-nucleic acid and in regions surrounding metal ions in protein structures. The clusters which occur in each of these regions differ in size, in the composition of residues and in the orientation of side-chains. Different orientations of side-chains will have different extent of side-chain interaction and therefore different values of $INT(R_i, R_j)$. Hence we have used different side-chain overlap (interaction) criteria measured by $INT(R_i, R_j)$ to detect clusters of various types. The classification based on the side-chain overlap is given below and the discussions in the subsequent sections are based on this classification.

Table 1. Normalization values for the 20 residue types derived from 148 protein structures

Residue type	Norm
Ala	55.7551
Arg	93.7891
Asn	73.4097
Asp	75.1507
Cys	54.9528
Gln	78.1301
Glu	78.8288
Gly	47.3129
His	83.7357
Ile	67.9452
Leu	72.2517
Lys	69.6096
Met	69.2569
Phe	93.3082
Pro	51.331
Ser	61.3946
Thr	63.7075
Trp	106.703
Tyr	100.719
Val	62.3673

The normalization values are obtained as mentioned in the text.

Side-chain clusters

Interactions between all side-chains (both polar and non-polar) are taken into account in constructing the graph. Further, a variable definition for the % side-chain overlap is used. (1) If the percentage interaction between two side-chains is 8 % or more, then it is defined as high side-chain overlap. (2) If the percentage interaction is greater than 5 % and less than 8 % then it is defined as a medium side-chain overlap. (3) If the percentage interaction is less than 5 % then it is classified as low side-chain overlap.

Hydrophobic clusters

In order to detect hydrophobic clusters, only the hydrophobic residues (L, I, M, V, P, F, C, A, Y, W) were considered in the protein. Two hydrophobic residues satisfying the percentage interaction criteria were connected in the graph. An overlap (extent of interaction) of about 8 % was too high to detect any cluster in this case and therefore a low percentage overlap criteria was used in detecting hydrophobic clusters. The percentage overlap between two hydrophobic side-chains in general is less than that of interacting exposed side-chains since the hydrophobic residues are usually found buried and surrounded by many other hydrophobic residues. However, hydrophobic residues found on the surface are surrounded by a relatively few hydrophobic residues. Therefore, along with the overlap criteria, a further classification based on the contact of the hydrophobic residues becomes important. A schematic representation given in Figure 1(c) shows that side-chain 1 is surrounded by four other side-chains to form a high contact hydrophobic network, and Figure 1(d) is a schematic of side-chain 1 having a low contact (surrounded by two side-chains), to form a low contact hydrophobic network. The classification used for the % side-chain overlap for detecting hydrophobic clusters was as follows:

(1) Two hydrophobic residues having a side-chain interaction greater than 5 % was classified as high hydrophobic overlap. (2) Interactions between 2 and 5 % was classified as medium hydrophobic overlap. (3) Less than 2 % was classified as low hydrophobic overlap. The classification based on the number of contacts was as follows:

(4) Hydrophobic residues having contact with more than three residues and satisfying a low side-chain overlap criteria with each of the side-chains in contact was classified as high contact

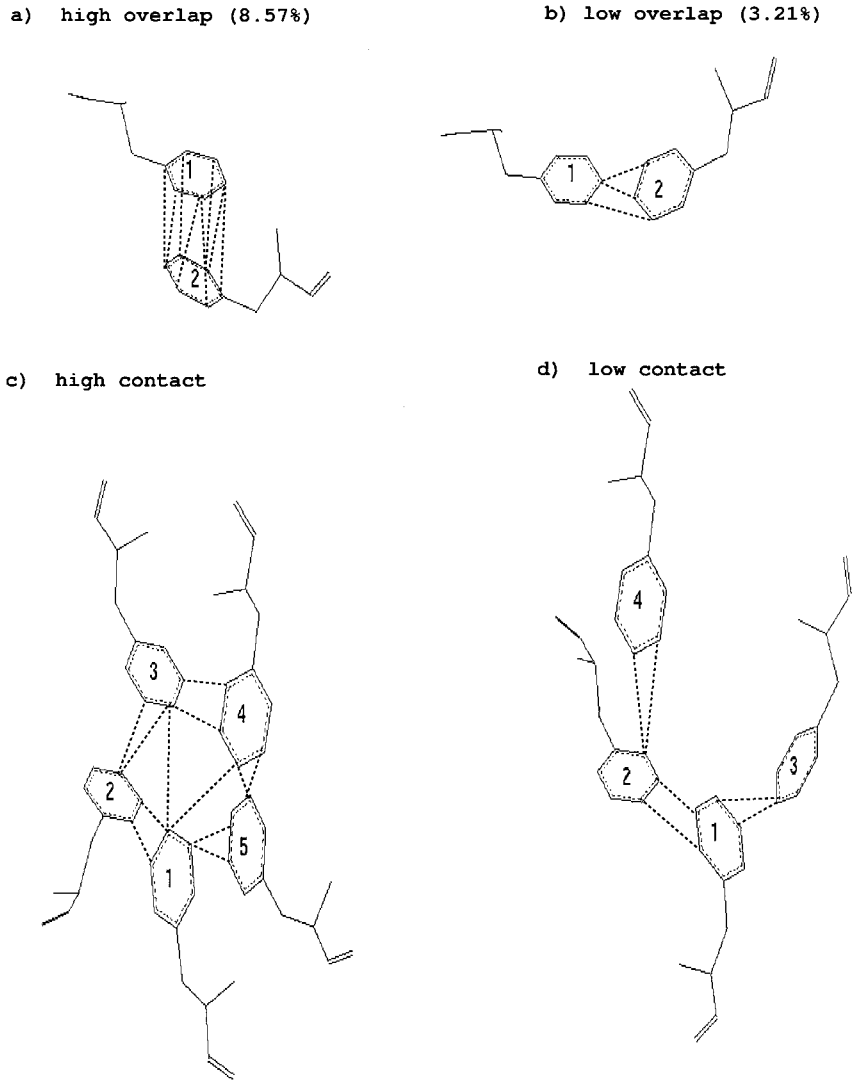

Figure 1. A schematic representation of side-chain overlap and contacts (as defined in the text). The dotted lines represent the interaction between side-chains.

hydrophobic. Only the high contact residues and the residues in contact were considered in constructing the graph. (5) Hydrophobic residues with at least two residues in contact and satisfying a low side-chain overlap criteria with the residues in contact was classified as low contact hydrophobic.

The above classification based on the side-chain overlap criteria for detecting side-chain clusters and hydrophobic clusters was arrived at after having experimented with a large number of side-chain overlap criteria ranging from 1% to 25% in intervals of 0.5% on a number of proteins. The number of clusters and the size of clusters detected vary considerably if extreme overlap criteria are used. Using 1% to detect high overlap side-chain clusters results in one big cluster which constitutes almost all the residues of the protein molecule, and using 8% results in four to five clusters of size three to six residues. However, if any value within the above classified range is used, then the resulting clusters do not differ significantly.

Detecting side-chain clusters in protein structures

In this section we illustrate with an example of T4 lysozyme(1LZ1) as to how the clusters and cluster centers were obtained from the eigenvalues and eigenvectors of the Laplacian matrix. The graph for the protein molecule lysozyme is constructed by considering the C^{β} atoms as vertices. Two C^{β} atoms were connected with an edge weight corresponding to their distance in space if the threshold side-chain criteria between the two side-chains was satisfied and the degree (number of other side-chains interacting) of at least one of the two side-chains under consideration was greater than one. If n number of side-chains in the protein satisfied this criteria then the graph would constitute n vertices. This way of constructing the graph eliminates two residue clusters and allows us to focus on clusters of significant size. The constructed graph is represented by a Laplacian matrix of size $n \times n$ and diagonalized to obtain the eigenvalues and

eigenvectors. Using a very high threshold (for example 25 %) does not give rise to any cluster and using a very low threshold (say 2 %) resulted in large expanded clusters.

The clusters obtained by using a threshold of 8.5 % in the protein lysozyme (1LZ1) is shown graphically in Figure 2(a) and in a tabular form in Table 2A. Column 4 corresponds to the sorted vector components of the second lowest eigenvalue, and the residues Tyr20, Lys97, Arg101 having the constant value of −0.310 form a three-residue cluster. In column 5 and 6 of Table 2A are tabulated the accessible surface area (Connolly, 1993) of the side-chains forming the cluster and the secondary structure (Kabsch & Sander, 1983) to which they belong. The absolute values of the vector components of the top eigenvalues are given in column 7 to 13. It is evident from column 7 to 13 that the vector components of the top eigenvalues have information on only one of the clusters, as the vector components are 0 for the residues which are not part of the cluster. Moreover, the magnitudes of the vector components of the top eigenvalues are correlated to the branching of side-chain (number of other side-chains interacting). For example in cluster 1, Tyr20 is highly

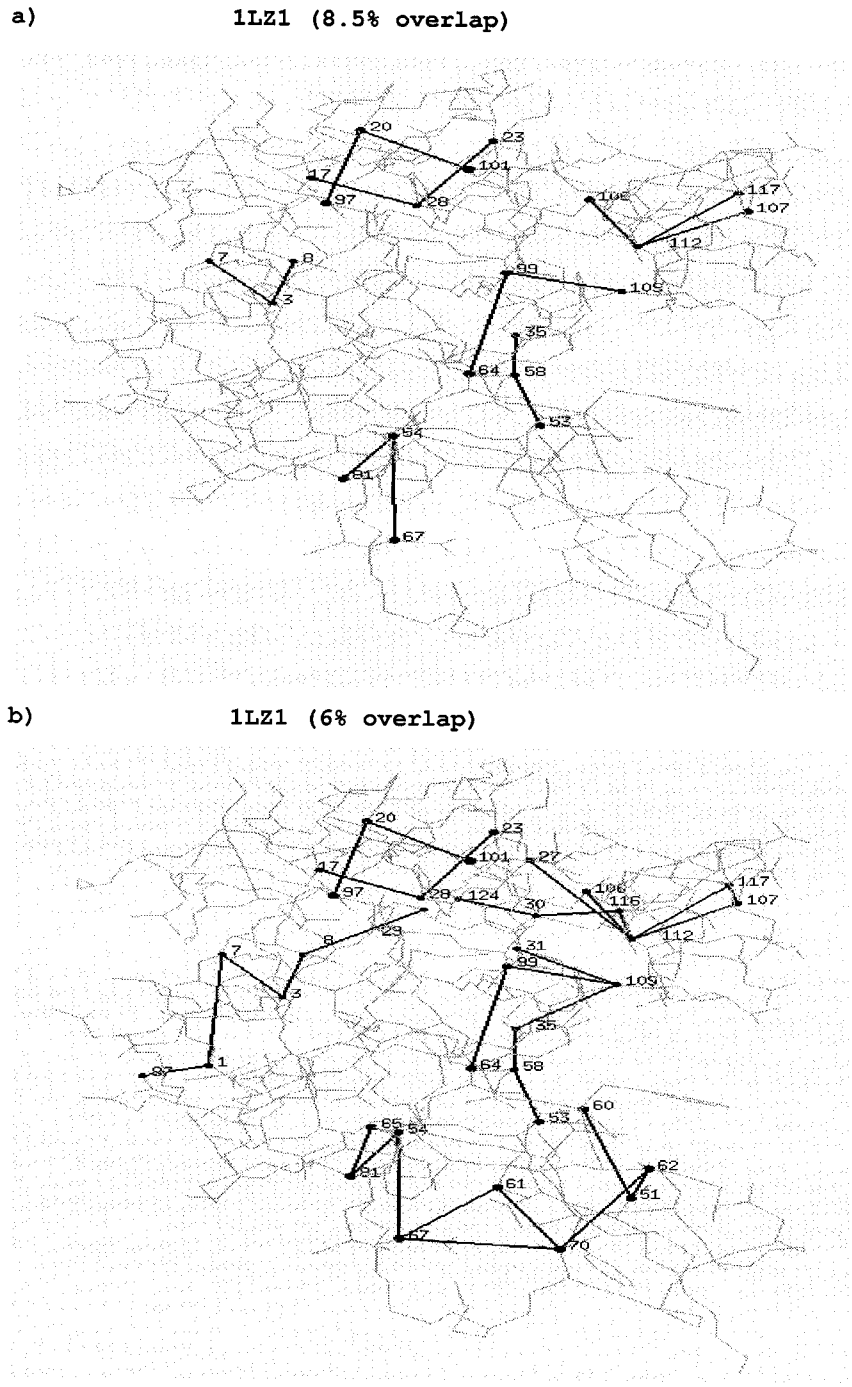

Figure 2. A connected graph representation of the detected clusters indicating the residue numbers (vertices) and their branching information. The same information is provided in Table 2A, Table 2B and Cluster Plots (Figure 3). The lysozyme molecule (1LZ1) is shown in LINE representation (Humphrey *et al.*, 1996).

Table 2. Clusters and eigenvector components in 1LZ1

A. *8.5% Overlap criteria*

Cl[a] no	Res. no[b]	Residue name	Eigenvector of 2nd lowest eigen value	% ASA[c]	SS[d]	Magnitude of the vector components of top eigenvalues 1	2	3	4	5	6	7
1	20	Tyr	−0.310	25.739	S2	0.000	0.000	0.000	0.000	**0.814**	0.000	0.000
	97	Lys	−0.310	24.702	H4	0.000	0.000	0.000	0.000	**0.354**	0.000	0.000
	101	Arg	−0.310	50.130	T11	0.000	0.000	0.000	0.000	**0.461**	0.000	0.000
2	3	Phe	−0.272	2.671	C2	0.000	0.000	0.000	0.000	0.000	**0.814**	0.000
	7	Glu	−0.272	37.944	H1	0.000	0.000	0.000	0.000	0.000	**0.466**	0.000
	8	Leu	−0.272	0.000	H1	0.000	0.000	0.000	0.000	0.000	**0.348**	0.000
3	54	Tyr	−0.024	10.498	S6	**0.795**	0.000	0.000	0.000	0.000	0.000	0.000
	67	Asp	−0.024	5.018	C7	**0.560**	0.000	0.000	0.000	0.000	0.000	0.000
	81	Cys	−0.024	1.372	H3	**0.235**	0.000	0.000	0.000	0.000	0.000	0.000
4	112	Trp	0.022	6.819	H6	0.000	**0.861**	0.000	0.000	0.000	0.000	0.000
	106	Ile	0.022	2.796	H5	0.000	**0.210**	0.000	0.000	0.000	0.000	0.000
	107	Arg	0.022	51.161	H5	0.000	**0.362**	0.000	0.000	0.000	0.000	0.000
	117	Gln	0.022	33.679	T13	0.000	**0.289**	0.000	0.000	0.000	0.000	0.000
5	28	Trp	0.030	0.000	H2	0.000	0.000	**0.816**	0.000	0.000	0.000	0.000
	17	Met	0.030	0.000	C3	0.000	0.000	**0.434**	0.000	0.000	0.000	0.000
	23	Ile	0.030	7.904	S3	0.000	0.000	**0.382**	0.000	0.000	0.000	0.000
6	99	Val	0.199	1.493	H4	0.000	0.000	0.000	**0.817**	0.000	0.000	0.000
	64	Trp	0.199	14.441	T8	0.000	0.000	0.000	**0.415**	0.000	0.000	0.000
	109	Trp	0.199	7.465	C12	0.000	0.000	0.000	**0.402**	0.000	0.000	0.000
7	58	Gln	0.349	3.191	T7	0.000	0.000	0.000	0.000	0.000	0.000	**0.811**
	35	Glu	0.349	16.328	H2	0.000	0.000	0.000	0.000	0.000	0.000	**0.323**
	53	Asp	0.349	23.420	S6	0.000	0.000	0.000	0.000	0.000	0.000	**0.488**

B. *6% Overlap criteria*

Cl[a] no	Res. no[b]	Residue name	Eigenvector of 2nd lowest eigen value	% ASA[c]	SS[d]	Magnitude of the vector components of top eigenvalues 2	3	7	8	9	11
1	20	Tyr	0.176	25.739	S2	0.000	0.000	0.000	0.000	0.000	**0.814**
	101	Arg	0.176	50.130	T11	0.000	0.000	0.000	0.000	0.000	**0.461**
	97	Lys	0.176	24.702	H4	0.000	0.000	0.000	0.000	0.000	**0.354**
2	7	Glu	0.221	37.944	H1	0.000	0.000	**0.709**	0.000	0.000	0.000
	1	Lys	0.221	35.378	C1	0.000	0.000	**0.631**	0.000	0.000	0.000
	3	Phe	0.221	2.671	C2	0.000	0.000	**0.269**	0.000	0.000	0.000
	87	Asp	0.221	58.695	C10	0.000	0.000	**0.128**	0.000	0.000	0.000
	8	Leu	0.221	0.000	H1	0.000	0.000	**0.098**	0.000	0.000	0.000
	29	Met	0.221	0.000	H2	0.000	0.000	**0.035**	0.000	0.000	0.000
3	67	Asp	−0.007	5.018	C7	0.000	**0.655**	0.000	0.000	0.000	0.000
	54	Tyr	−0.007	10.498	S6	0.000	**0.509**	0.000	0.000	0.000	0.000
	81	Cys	−0.007	1.372	H3	0.000	**0.420**	0.000	0.000	0.000	0.000
	65	Cys	−0.007	0.000	C6	0.000	**0.264**	0.000	0.000	0.000	0.000
	61	Ser	−0.007	0.000	T8	0.000	**0.203**	0.000	0.000	0.000	0.000
	70	Thr	−0.007	0.021	C7	0.000	**0.144**	0.000	0.000	0.000	0.000
	62	Arg	−0.007	15.402	T8	0.000	**0.059**	0.000	0.000	0.000	0.000
	51	Ser	−0.007	0.464	S6	0.000	**0.019**	0.000	0.000	0.000	0.000
	60	Asn	−0.007	21.356	S7	0.000	**0.007**	0.000	0.000	0.000	0.000
4	112	Trp	0.030	6.819	H6	**0.806**	0.000	0.000	0.000	0.000	0.000
	30	Cys	0.030	0.000	H2	**0.399**	0.000	0.000	0.000	0.000	0.000
	107	Arg	0.030	51.161	H5	**0.251**	0.000	0.000	0.000	0.000	0.000
	116	Cys	0.030	0.028	T13	**0.184**	0.000	0.000	0.000	0.000	0.000
	117	Gln	0.030	33.679	T13	**0.182**	0.000	0.000	0.000	0.000	0.000
	27	Asn	0.030	26.481	H2	**0.148**	0.000	0.000	0.000	0.000	0.000
	106	Ile	0.030	2.796	H5	**0.144**	0.000	0.000	0.000	0.000	0.000
	124	Tyr	0.030	9.776	H7	**0.135**	0.000	0.000	0.000	0.000	0.000
5	28	TRP	−0.416	0.000	H2	0.000	0.000	0.000	0.000	**0.816**	0.000
	17	Met	−0.416	0.000	C3	0.000	0.000	0.000	0.000	**0.434**	0.000
	23	Ile	−0.416	7.904	S3	0.000	0.000	0.000	0.000	**0.382**	0.000
6	109	Trp	−0.112	7.465	C12	0.000	0.000	0.000	**0.764**	0.000	0.000
	99	Val	−0.112	1.493	H4	0.000	0.000	0.000	**0.516**	0.000	0.000
	31	Leu	−0.112	0.000	H2	0.000	0.000	0.000	**0.291**	0.000	0.000
	64	Trp	−0.112	14.441	T8	0.000	0.000	0.000	**0.185**	0.000	0.000
	35	Glu	−0.112	16.328	H2	0.000	0.000	0.000	**0.171**	0.000	0.000
	58	Gln	−0.112	3.191	T7	0.000	0.000	0.000	**0.038**	0.000	0.000
	53	Asp	−0.112	23.420	S6	0.000	0.000	0.000	**0.009**	0.000	0.000

[a] Cl no, Cluster number.
[b] Res. no, residue number.
[c] ASA, accessible surface area.
[d] SS, secondary structure (S, sheet; H, helix; T, turn; C, coil).

branched (interacts with two other side-chains) as compared to Lys97 and Arg101 (Figure 2(a)) and has the highest vector component magnitude of 0.814 (Table 2A (column 11)). Thus, a highly branched side-chain or the side-chain which is connected to highly branched side-chains, has the largest absolute vector component value in the corresponding top eigenvalue in the cluster. Such a side-chain (vertex) is defined to be the center of the cluster. Hence branching information is a byproduct of the top eigenvalue (Randic, 1975). Also in the other clusters, the residue forming center of the cluster has the largest vector component corresponding to the top eigenvalues. Thus to summarize, the vector component of the second lowest eigenvalue gives information on the clustering of side-chains and the vector components of the top eigenvalues give information on its branching (number of interactions with other side-chains). The detected clusters are shown graphically in the form of a two-dimensional plot called the Cluster Plot (Figure 3(a)). Information on the number of clusters detected, the residues forming the clusters, the center of clusters and the solvent accessibility are obtained by just visualizing the Cluster Plot. The first residue in each cluster of the cluster plot corresponds to the center of the cluster. For example Tyr20, Phe3, Tyr54, Trp112, Trp28, Val99 and Gln58 form the centers of seven clusters (Figure 3(a)).

Expanded cluster: network of side-chain interaction

Since the residues are connected based on % overlap between the side-chains, the detected clusters are sensitive to the overlap criteria used. As the percentage side-chain overlap cutoff is reduced, some of the clusters are found to expand by a network of additional side-chain interactions. A threshold of 6% is used on the same protein (1LZ1) to show how the previously detected clusters expand. Column 4 in Table 2B shows that six clusters are detected using a reduced cutoff and a comparison with Table 2A shows that cluster 6 is formed by merging of clusters 6 and 7 observed using a high percentage cutoff. This is also shown in Figure 2(b). Clusters 2, 3 and 4 expand by adding in extra residues to the cluster. However clusters 1 and 5 remain the same in spite of using a reduced cutoff criteria. The rapid expansion of one or two clusters as the percentage overlap criteria is reduced is a general feature observed in all the proteins studied and its probable significance to protein folding and structure is discussed in a later section.

The center of the cluster is also found to shift as the cluster expands. Using a high % overlap criteria Phe3 forms the center of cluster 2 (Table 2A) and gets shifted to Glu7 in a reduced overlap criteria (Table 2B). Cluster 6 is formed by the merging of clusters 6 and 7 observed using a 8.5% overlap criteria. The cluster center shifts to Trp109 in cluster 6 (Table 2B) as this residue links the two clusters 6 and 7 (Figure 2(b)) observed using 8.5% overlap criteria.

Materials

The protein coordinates and the entry names used are obtained from the Brookhaven database (Bernstein *et al.*, 1977). The program is written in C, Fortran and MATLAB. The program is interactive and has options for side-chain overlap criteria to be defined by the user. The flowchart of the program is given below (Figure 4). Although the present study is restricted to hydrophobic and all residue side-chain clusters, the program can also be used for detecting exclusively charged clusters as shown in the flowchart. Interested users can mail their queries to sv@mbu.iisc.ernet.in.

Results and Discussion

As there is no uniformity in the structure of proteins in terms of parameters such as packing density, non-bonded contacts and pairwise side-chain interaction (Richards, 1974; Beardsley & Kauzmann, 1996), any analysis based on rigid criteria can miss some important features. For example a commonly used approach to assign a residue to the hydrophobic cluster or nucleus is based on the accessibility information of the residue (Plochocka *et al.*, 1988). This information is necessary but not sufficient to study the topology, structure and other properties of non-polar regions of the protein, as hydrophobic clusters are known to occur also on surfaces (Laurence & Evans, 1995).

In order to overcome this problem we have in the present analysis constructed graphs based on several criteria (see Algorithm) which are, (1) high side-chain overlap, (2) medium side-chain overlap, (3) low side-chain overlap, (4) high hydrophobic overlap, (5) medium hydrophobic overlap, (6) low hydrophobic overlap, (7) high contact hydrophobic and (8) low contact hydrophobic. The results of the analysis are presented below.

High overlap side-chain clusters

Residue preferences

An 8.5% side-chain overlap criteria was used to detect high overlap side-chain clusters. All the residues in the protein were included for the clustering procedure. The clusters detected using a high side-chain overlap criteria on six proteins myoglobin (4MBN), hemoglobin (2LHB), ribonuclease A(7RSA), angiogenin (1AGI), hen egg white lysozyme (1LZC) and alpha lactalbumin (1ALC) are shown in the form of cluster plots in Figure 5. The detected clusters are dominated by charged residues like K, E and aromatic residues like F, H, Y and W. A similar trend in the preference of residues in clusters was observed for a larger dataset of proteins studied by Heringa & Argos (1991).

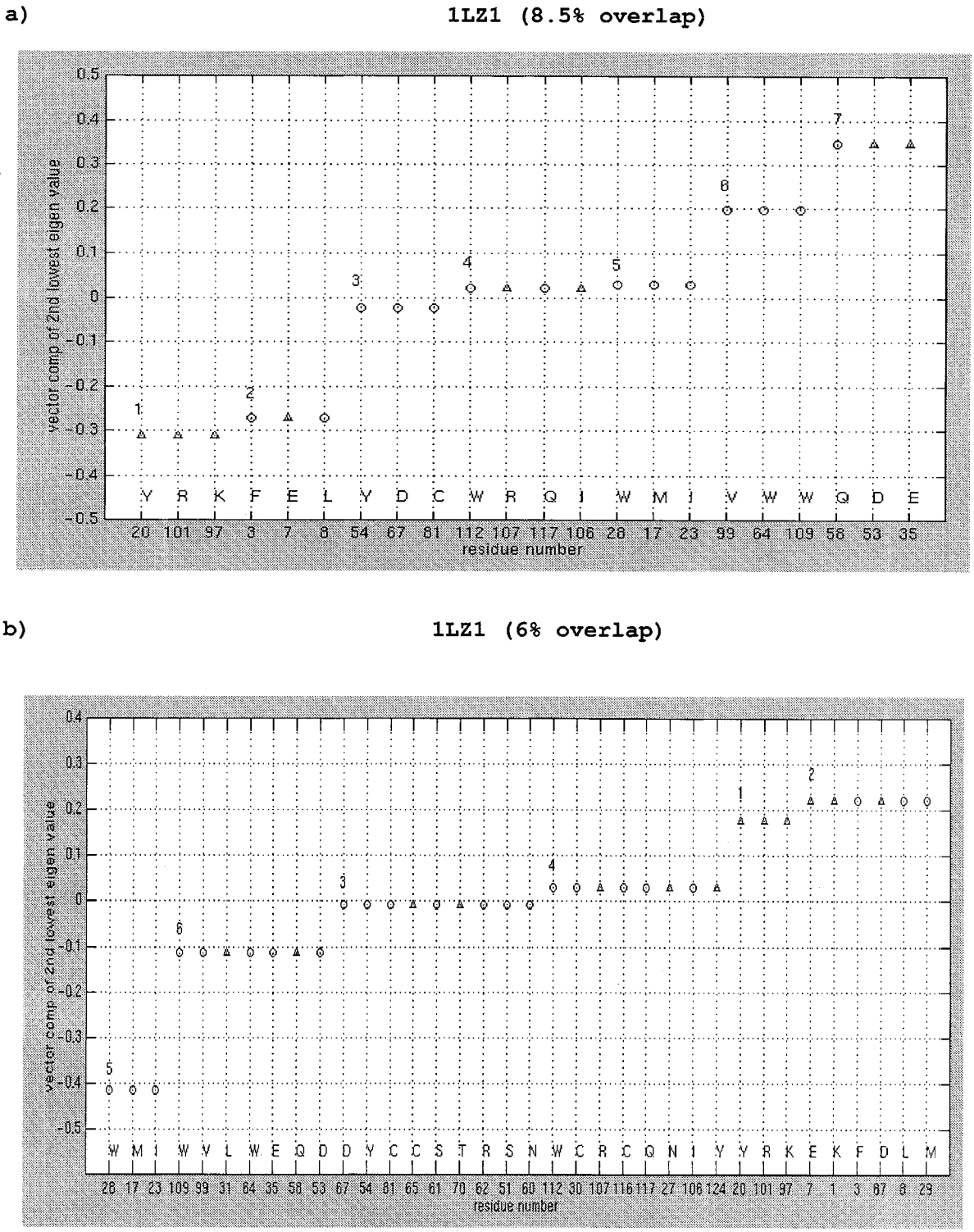

Figure 3. Cluster Plot: the *X*-axis denotes the residue number and the residue type (single letter code of amino acid) is shown above the *X*-axis. *Y*-axis denotes the vector components of second lowest eigenvalue. The symbols (⊙) denotes buried, (△) partially exposed and (⋆) completely exposed side-chains. The first residue is the center of the cluster (cluster number marked above the residue).

Cluster location

The solvent accessibility of the side-chains forming a cluster were calculated. The side-chains having solvent accessibility less than 15% were considered buried, accessibility between 15 to 60% were considered partially exposed and accessibility greater than 60% were considered highly exposed. This is denoted by different symbols in the cluster plot (Figure 5). It is clear from Figure 5 that most clusters have partially exposed side-chains, and are

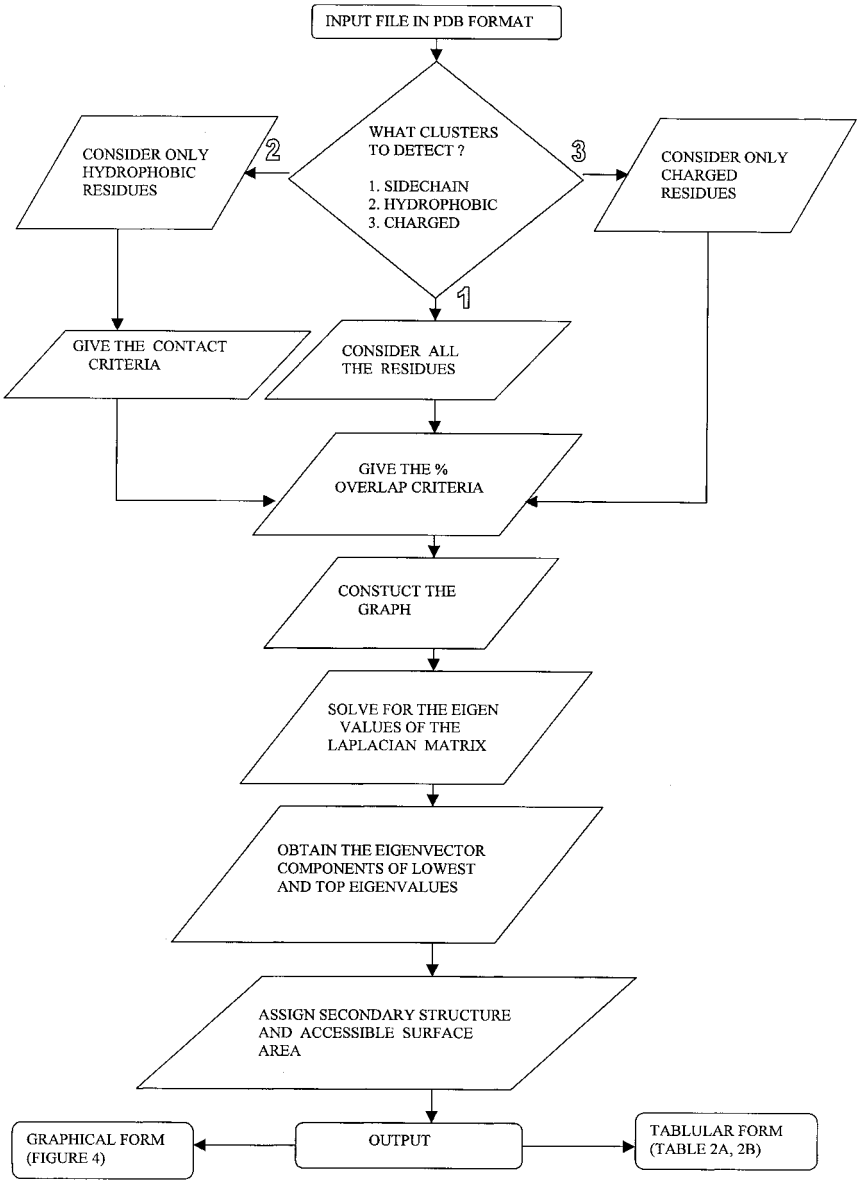

Figure 4. Flow chart of the algorithm.

located on the protein surface. This trend is also observed in a larger data set of proteins (Heringa & Argos, 1991). It is interesting to note the presence of buried, polar and charged side-chains in the clusters. For example, D66 of cluster 1 in hen egg white lysozyme (1LZC) is a buried charged residue and N57 in cluster 2 of alpha lactalbumin (1ALC) is a buried polar side-chain. Cluster 1 of myoglobin (4MBN) has two buried histidine residues H24 and H119. Detecting such buried charged residues in the clusters would be of interest as charge or polar groups play an important role in generating unique structures and conferring conformational stability (Fersht, 1984; Harbury *et al*., 1993; Honig & Yang, 1995). Also, buried polar interactions are known to reduce the rate of folding (Waldburger *et al*., 1996). Charged residues in the cluster can also be possible targets of mutation to study the role of residues in the stability and folding of proteins.

Thermal factors

The flexibility of protein atoms in a crystal structure is measured by its thermal factors called the *B* factor (Stout & Jensen, 1968). The lower the *B* factor, the more rigid is the atom associated with it. The average *B* factor of a protein was calculated by averaging over the *B* factors of all the atoms obtained from protein crystal structures. The *B* factor of the cluster was determined by averaging over all the atoms of the residues which form the cluster (Table 3). It is observed that the average *B* factors of the clusters are more often close to the average *B* factor of the protein. Although the average *B* factors are generally correlated with the exposure of the residue, there is no strict corre-

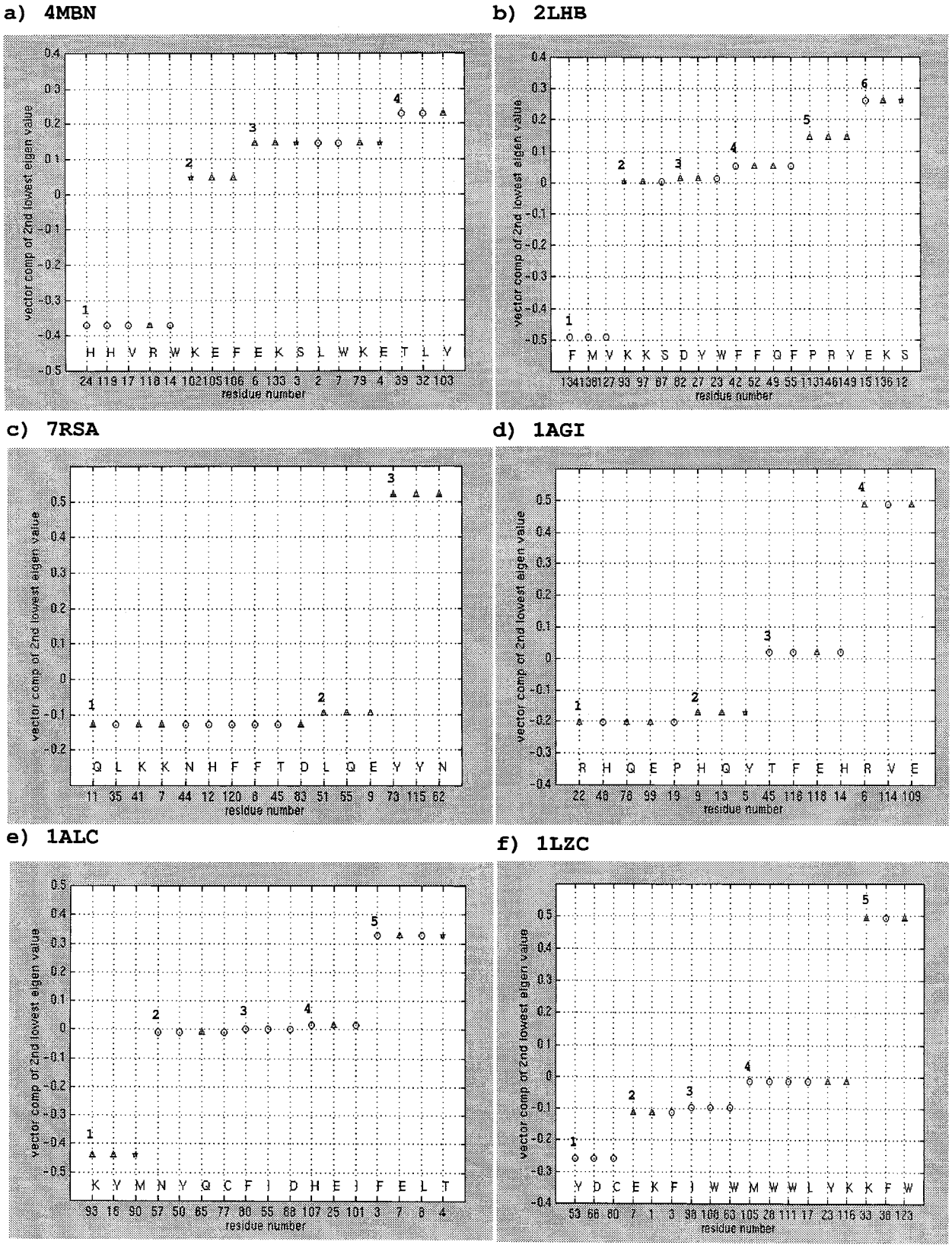

Figure 5. Cluster Plots. High overlap side-chain clusters (a)-(f). Myoglobin (4MBN), hemoglobin (2LHB), ribonuclease A(7RSA), angiogenin(1AGI), hen egg lysozyme(1LZC), alpha lactalbumin(1ALC). The axis and symbols represent the same as in Figure 3.

Table 3. Thermal factors for the detected clusters

PDB code	Cluster number (active site cluster in bold)	% Exposed residues in the cluster	Average *B* factor of the cluster	Average *B* factor of the protein
7RSA	**1**	**40**	**14.3**	15.4
	2	100	14.7	
	3	100	16.5	
1AGI	1	60	31.2	29.51
	2	100	20.3	
	3	**25**	**9.2**	
	4	66.6	14.8	
4MBN	1	40	13.1	13.59
	2	100	12.7	
	3	71.4	13.8	
	4	**33.3**	**12.2**	
2LHB	1	0	12.6	17.75
	2	66.6	23.4	
	3	66.6	12.4	
	4	**50**	**15.2**	
	5	100	21.5	
	6	100	20.5	
1ALC	1	100	30.9	29.8
	2	**25**	**16.3**	
	3	0	23.9	
	4	33.3	28.5	
	5	50	28.9	
1LZC	1	0	10.83	16.48
	2	66.6	17.10	
	3	**0**	**8.44**	
	4	33.3	11.30	
	5	66.6	16.23	

spondence as seen from Table 3. Further the clusters marked in bold are the clusters found near the active site and are found to have the lowest *B* factors among the clusters detected, except in the case of 2LHB.

Stability

The majority of the residues forming clusters emanate from different secondary structural elements of the protein, stabilizing the tertiary fold. For example, cluster 4 in alpha lactalbumin (1ALC) (Figure 5) is formed by H107, E25 and I101 which are from three different helices. This cluster should be important for the stability of this fold since an equivalent cluster (cluster 4) constituting residues M105, W28, W111, L17, Y23 and K116 is detected in a structurally equivalent position in hen egg white lysozyme (1LZC) (Figure 5) which has the same fold as that of alpha lactalbumin. In the case of myoglobin (4MBN), cluster 4 is formed by residues T39, L32 and 103Y and the residues are from helices B, C and G, respectively. However, no equivalent cluster was detected in 2LHB as the packing of helices in 2LHB is different (Weaver, 1992). Cluster 3 in hemoglobin (2LHB) is formed by D82, Y27 and W23. Y27 and W23 from the middle of helix A and D82 from the middle of helix E stabilize the orientation of the two helices.

Active site: expanded clusters

Generally one of the clusters in the high overlap side-chain cluster is close to the active/binding site of the proteins studied (Table 4). For example cluster 4 in myoglobin (4MBN) occurs near the binding site constituting residues 39T, 32L and 103Y (Figure 6(a)) of which 103Y interacts with the porphyrin ring (Takano, 1977). As mentioned earlier this cluster also has a low *B* factor (Table 3). The most interesting feature of the cluster close to the active site is the way in which they expand as the side-chain overlap criteria is reduced (low overlap) to form a contiguous network. Only one or two clusters expand when the overlap criteria is reduced and invariably the active site cluster is one of them (Table 4).

The active site clusters detected with high overlap criteria and low overlap criteria for three proteins; myoglobin (4MBN), ribonuclease A (7RSA) and hen egg white lysozyme (1LZC) are shown in Figure 6. As the overlap criteria is reduced from high side-chain overlap to low side-chain overlap, a few clusters expand by incorporating other residues into the cluster. This expansion of a few clusters may be important from the folding point of view as the nucleation condensation hypothesis supports the formation of an expanded cluster around the nucleation center (Fersht, 1997). It has been shown by protein engineering methods that the residues which are involved in the catalysis are delocalized over the whole active site and are not restricted to just a few key residues (Fersht, 1987). The expansion of particularly the active site cluster is significant from this perspective. The expanded clusters may also have implications at the structural level as the expanded network of side-chain interactions surrounding the active site (Figure 6)

Table 4. Active site clusters and expanding clusters

PDB code	Active site cluster number[a]	Residues of the cluster near the active/binding site[b]	Expanding cluster number[c]
1ALC	2	Y50 **N57** C77 Q65	**2**
1LZC	3	I98 **W63** W108	**3**
7RSA	1	**Q11** K7 L35 **K41 H12** F8 **F120 N44 T45** D83	**1**
1AGI	3	**T45 F116 E118 H14**	**3** 1
4MBN	4	T39 L32 **Y103**	**4** 1
2LHB	4	F42 Q49 **F52** F55	**4** 1 5 6

[a] Details are given in Table 3.
[b] The residues interacting with the ligand are shown in bold.
[c] The active site cluster number is shown in bold.

may be important in aiding the active site cleft movement during the process of ligand binding.

Hydrophobic clusters

Structural similarity

It is a known fact in protein structures that the majority of non-polar side-chains pack together to form the core of the protein. Among sequences which take the same topology, the buried hydrophobic residues are known to be the most conserved (Murzin *et al.*, 1992; Bashford *et al.*, 1987). Also, the conserved residues are shown to be important in the mechanism of folding (Shakhnovich *et al.*, 1996). In order to assess the efficacy of our algorithm, we have applied our method on topologically similar proteins and have shown that the residues which form the clusters are conserved in the two proteins.

The hydrophobic clusters were detected using a medium hydrophobic overlap criteria (2.5 %). The same % overlap criteria was used on the two topologically similar proteins. The detected clusters are shown to occupy structurally similar locations among the pairs of proteins considered (Figure 7). The residues forming clusters are mostly buried. Even though there are a few exposed hydrophobic clusters detected, they are found to occupy an equivalent position in the two topologically similar proteins. Ribonuclease A (7RSA) and angiogenin (1AGI) are proteins with the same fold with a sequence similarity of 33.3 %. The detected clusters are shown to occupy structurally equivalent positions among the two proteins. Most residues forming the cluster in the two proteins are also conserved in the two sequences as shown by a symbol (*) in Figure 7. Similarly, in the case of myoglobin (4MBN) and hemoglobin (2LHB) the conserved hydrophobic residues and their locations in the structure are shown in Figure 7. The same observation also holds for alpha lactalbumin and lysozyme. Hence, the algorithm is robust enough to detect such conserved features in structurally similar proteins and can serve as a useful tool in protein structure analysis.

Folding intermediates

A lot of experimental work has gone into identifying early folding intermediates. Hydrogen exchange experiments are used to monitor the exchange rates of the backbone amides with the bulk solvent. A reduction in the exchange rate of the amide protons is due to its protection from the bulk solvent by the protein atoms during the process of folding (Chyan *et al.*, 1993; Lu & Dahlquist, 1992).

The detected hydrophobic clusters with low hydrophobic overlap criteria (1 %) show a good correlation with the protection factors. The residues belonging to various clusters are listed in Table 5. The residues which get protected are shown in bold. The protection of amide hydrogen atoms during the refolding of hen egg white lysozyme was measured by Radford *et al.* (1992). The amide protons of the residues which get protected within 9 ms of the refolding process are shown in bold in Table 5. It is observed that most of the residues in cluster 1 get protected within 9 ms. The amide proton of at least one residue is found to get protected within 6 ms of the folding process in all the five clusters (Table 5) detected in myoglobin (Jennings & Wright, 1993) emphasizing the fact that the detected clusters are very important from the folding perspective. The degree of protection of amide protons on RNase A has been studied (Udgaonkar & Baldwin, 1990). The amide protons were classified based on the degree of protection, namely strong protection, moderate protection and weak protection. The residues of the detected clusters which show strong degree of protection are shown in bold in Table 5. For example V54, V47, I81, I106 and V108 of cluster 1 show a strong degree of protection. Similarly V63 in cluster 2 is measured to be strongly protected but Y97 in cluster 3 shows a moderate protection. Thus many of the residues detected in the hydrophobic clusters appear to have participated during the intermediate stages of protein folding and seem to be important from the folding point of view.

The emphasis so far in the literature has been in correlating folding intermediates with hydrophobic regions (Evans *et al.*, 1991; Pen & Briggs, 1992;

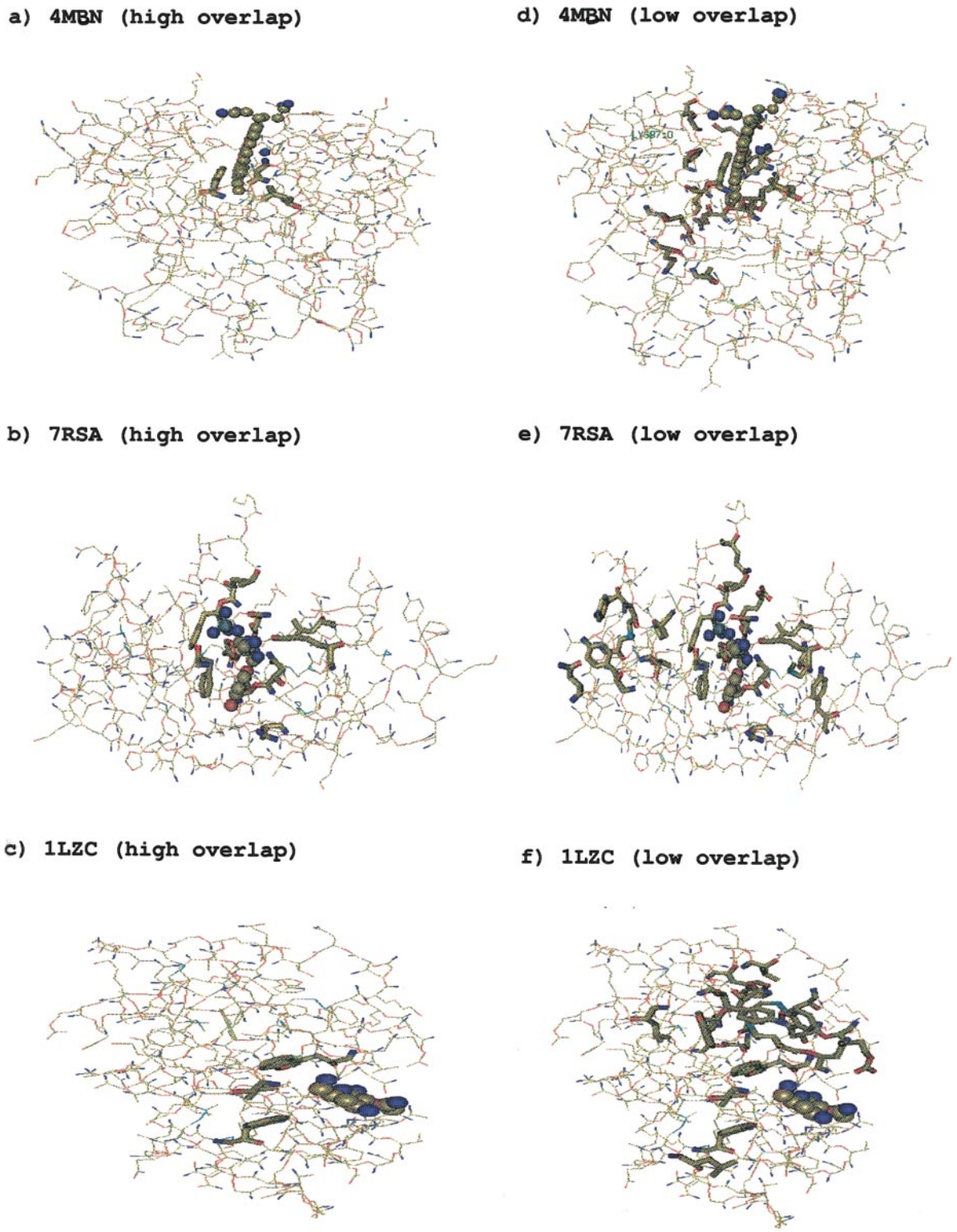

Figure 6. The active side clusters of 4MBN, 7RSA and 1LZC. Figures on the left ((a), (b), (c)) are the active site clusters. Figures on the right ((d), (e), (f)) are the corresponding expanded clusters. The ligands are shown in VDW representation and cluster residues are shown in BONDS representation. Figures were generated using VMD (Humphrey *et al.*, 1996).

Gronenborn & Clore, 1994; Zehfus, 1995). However, polar interactions are shown to be barriers during the folding process (Waldburger *et al.*, 1996). High overlap side-chain clusters detected were examined to see if any charged residues in the cluster show high protection factors. Interestingly the charged/polar residues which get protected form the center of the cluster. For example

Table 5. Hydrophobic clusters

PDB code	Cluster number	Hydrophobic residues forming cluster[a]
1ALC	1	P24 L115 L119 I27 L23 Y36 F31 W118
	2	L26 I21 I15 L12 I85 L8 F3
	3	F80 I55 I75 M30 F53 **W104** A92 W60 I72 I95 **Y103**
	4	I41 **Y50 L81**
1LZC	1	F3 **L8** I55 I88 **M12** F38 **L17 W28** L56 **V92 Y23** M105 **W111 V99 W108 V29 W123** A32 I58 **W63** L83 I98 L75 Y53 **I78 A11** Y20 **F34**
	2	I124 L25 L129
4MBN	1	I28 I111 **V114** L135
	2	F123 **V13 L115** M131 **V10** W7 M131
	3	**F138** I75 **I142** L86 A94 Y146 Y151 I101
	4	**F33** L40 F43 M55 F46 L49 L61 **I30**
	5	**W14 V17** L72 L76
7RSA	1	M13 F8 L51 **V54 V47** M79 **I81** F120 **I106** A102 **V108** V57 P117
	2	I107 **V63** A122 V124
	3	M30 L35 F46 **Y97** M29 A20 Y25

[a] Experimentally identified residues as protected in the folding intermediate are marked in bold.

the amide proton of K33 in cluster 5 in alpha lactalbumin gets protected within the first minute of folding (Buck *et al.*, 1993) and is also the center of cluster 5 (Figure 5(b)). A similar correlation between the charged/polar residues having high protection factors and the residue forming the center of the cluster is observed in other proteins as well. For example, in the case of cluster 1 of ribonuclease A, Q11 forms the center of the cluster as seen from Figure 5(c). Q11, adjacent in sequence to the catalytically active H12 shows a weak degree of protection (Udgaonkar & Baldwin, 1990) and is also a conserved residue in the ribonuclease A family.

Hydrophobic clusters: Φ values

The recent advances in protein engineering and NMR procedures have contributed to our ideas on protein folding (Fersht, 1997). Using these methods, it has been possible to identify residues important in the transition state of folding by performing point mutations. The effect of the engineered mutations on the kinetics and the thermodynamics of the protein molecule is used as a probe to study the structure formation during the folding process. Each mutation alters specific side-chain interactions in the transition and the native state resulting in a difference of stability between the mutant and wild-type protein for the two states. This effect of the mutation is estimated by the Φ_F value which is obtained by the ratio $\Delta\Delta G_{T-D}/\Delta\Delta G_{N-D}$ (Fersht, 1993), where $\Delta\Delta G_{N-D}$ is the change in stability of the protein on mutation (N, native state; D, denatured state) and $\Delta\Delta G_{T-D}$ is the change in stability of the transition state of folding (T, transition state; D, denatured state). If Φ_F is 1, the transition state is disrupted by mutation by the same energy as is the fully folded protein and so indicates complete formation of native structure in the transition state, while Φ_F of 0 shows that the transition state is as insensitive to mutation as is the fully denatured state.

In the present analysis we have attempted to correlate the observed Φ_F values with the detected hydrophobic residues in the cluster. We find a good correlation between the vector components of the top eigenvalues and the reported Φ_F values. Detailed Φ value analysis has been done for chymotrypsin inhibitor (1CIQ) (Itzhaki *et al.*, 1995). Mutations have been done on the residues forming the core and mini core of the protein. Clusters 2 and 3 correspond to the core and mino-core, respectively (Figure 8(a)). The largest Φ_F of 0.53 was measured for a LA49 mutant in the core. Leu49 is found to be the center of the hydrophobic cluster detected (Figure 8). A correlation coefficient of 0.87 was obtained between Φ_F values observed in the core mutations and the vector components of the top eigenvalue of the hydrophobic cluster. A similar correlation was also obtained for the mini-core. The Φ_F values and the vector components are tabulated in Table 6A.

Detailed Φ value analysis has also been done on barnase (1RNB) (Matouschek *et al.*, 1992). The refolding pathway comprises a folding intermediate, a major transition state and the fully folded structure. A classification based on the Φ value was done to quantify the effect of mutation on the three states observed during the folding process, namely the intermediate state, transition state and the final folded state. The Φ values obtained from the transition state were classified as Φ_T and from intermediate states were classified as Φ_I. For interactions which are same in the transition state and in the folded state $\Phi_I \leqslant \Phi_T \approx 1$. This is found for IV88 mutation which has the largest vector component magnitude of 0.763 in the hydrophobic cluster 3 (Table 6B) and forms the center of the cluster (Figure 8(b)). If the interactions in the transition state are stronger or more frequent than in the intermediate state then $\Phi_I \leqslant \Phi_T < 1$. This is observed for mutations IV109, VT10, IA76 and LA14. The magnitudes of the vector components are comparable in these cases. The two mutations whose Φ_T values do not correlate with the magnitude of the vector component are YA13 and YA17. A possible reason for this disagreement could be that the two tyrosine residues are highly exposed to the solvent in the folded state and could have undergone minor conformational changes after the final folded state was attained. On the other hand, the buried hydrophobic residues show a good correlation with the Φ_T values. For mutation IV96, $\Phi_I \approx \Phi_T < 1$, means this region is formed both in the transition state and intermediate state but less formed in the folded state. The magnitude of the

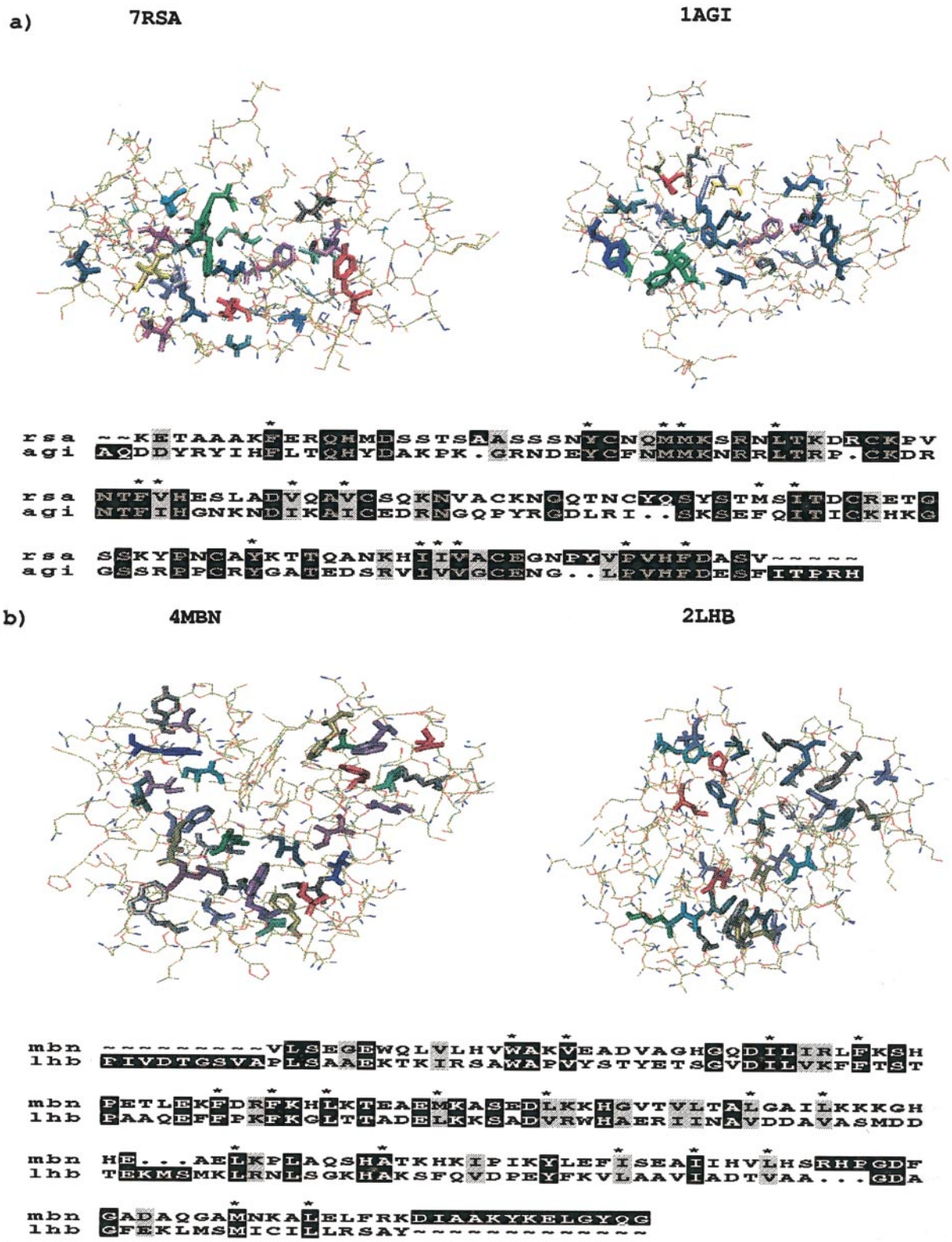

Figure 7. (a) Top: Structural location of high contact hydrophobic cluusters in topologically similar proteins 7RSA and 1AGI. Below: Sequence alignment. Identical residues are denoted by dark shading; equivalent residues by grey shade; residues forming the cluster by *. (b) Structurally similar proteins 4MBN and 2LHB. The description of top and bottom panels are the same as that of (a). The clusters are shown in BONDS representation and the protein in LINE representation.

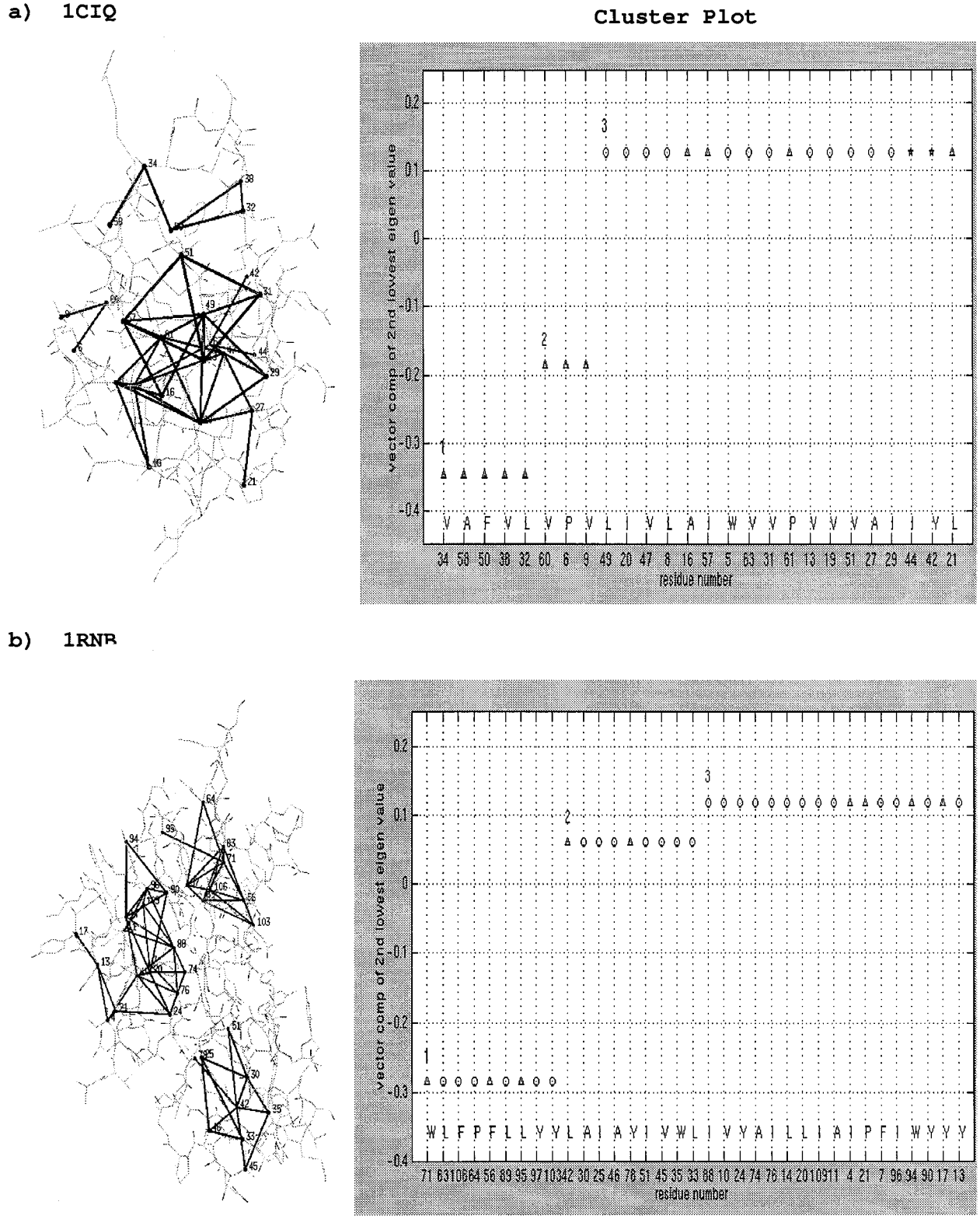

Figure 8. Left panel: Connected graph representation of high contact hydrophobic cluster in chymotrypsin inhibitor 2 (1CIQ) (Neira *et al.*, 1996) and barnase (1RNB) (Baudet & Janin, 1991). The numbered C^{β} atoms form the vertices and edges represent interacting hydrophobic side-chains. Right panel: The corresponding cluster plots. The center of the cluster L49 in (1CIQ) and I88 in (1RNB) are shown in the connected graph and in the corresponding cluster plot.

vector component corresponding to IV96 is very small (0.043) (Table 6B). $\Phi_I \approx \Phi_T \approx 0$ is observed for IV4 mutation which shows that the structure probed by the mutation is as unfolded in the intermediate and transition state as it is in the unfolded state. The corresponding vector component magni-

Table 6. Correlation of exprrimental Φ_F values with the vector components

A. *In 1CIQ*

Protein code	Mutation in the core	Φ_F	Magnitude of vector components of the top eigenvalue
1CIQ	**LA49**	**0.53**	**0.665**
	IV20	0.40	.627
	LA51	0.25	0.035
	IV29	0.17	0.029
	LA8	0.15	0.166
	PA61	0.02	0.078
	mini-core		
	FL50	0.28	0.478
	VA38	0.26	0.206
	LA32	0.21	0.088

B. *In 1RNB*

Protein code	Mutation	Φ_I	Φ_T	Magnitude of vector components of the top eigenvalue
1RNB	**IV88**	**0.7**	**0.9**	**0.763**
	IV109	0.4	0.6	0.174
	VT10	0.3	0.4	0.326
	IA76	0.2	0.5	0.213
	LA14	0.5	0.6	0.208
	IV96	0.6	0.6	0.043
	IV4	0.0	0.0	0.049
	YA13	0.4	0.5	0.000
	YA17	0.5	0.6	0.000

Φ_F, Φ_T and Φ_I values are described in the text.

tude is also the least and close to 0. The overall correlation coefficient of the vector components with Φ_T values measured in the transition state is 0.55.

The cluster plot can be useful in predicting mutations to study the transition state of folding. For instance, the residues Y24 and A74 have significantly high vector component magnitudes in cluster 3 of 1RNB, however their Φ_F values have not been reported. Hence we suggest that mutation of these residues could have a larger effect on the transition state of folding. The technique however is limited by the fact that one cannot predict on those residues which do not form part of the clusters detected.

Multi-domain proteins

With the increasing number of protein structures in the protein database, a structural classification is essential. A major challenge in such a classification is in assigning the domains. A database of structural domains has been presented (Sowdhamini *et al.*, 1996), where the domains were identified based on the clustering of secondary structural elements. As there is no universal definition of a domain, the other existing algorithms use different criteria for domain identification. Some of the algorithms reported (Holm & Sander, 1994; Siddiqui & Barton, 1995; Islam *et al.*, 1995) are based on the fact that the residues forming a domain would make more internal contacts within themselves than with the rest of the protein. The DOMAK algorithm (Siddiqui & Barton, 1995) calculates split values from the number of each type of contact when the protein is divided, the split value being larger when the two parts of the structure are different. A consensus approach for assignment of structural domains has been presented (Jones *et al.*, 1998) which uses all the available algorithms and assigns domains based on the consensus. For proteins which could not be done by consensus procedure, the domain boundaries were assigned using any one of the algorithms and then the domains were analyzed for trend in size and secondary structure. The criteria for our algorithm is similar to that of Swindells (1995) in which each domain in a protein is considered to have a hydrophobic core and detecting the cores would correspond to detecting the domains. We have used high hydrophobic contact criteria to construct the graph. The resulting clusters are compact, buried and form the core of the protein. By looking at the cluster plot, the number of cores and hence the number of domains can be deduced. Figure 9 shows the cores detected in glutathione reductase (3GRS) which has three domains and hydrolase elastase (1EZM) having two domains. When a low hydrophobic contact criteria is used, the hydrophobic cores start to expand by adding in additional hydrophobic residues (Figure 9). A few hydrophobic clusters on the surface and domain interface are also detected.

In the above sections we have demonstrated the applicability of our method in detecting clusters which are of interest from protein structure, function and folding points of view. Other interesting

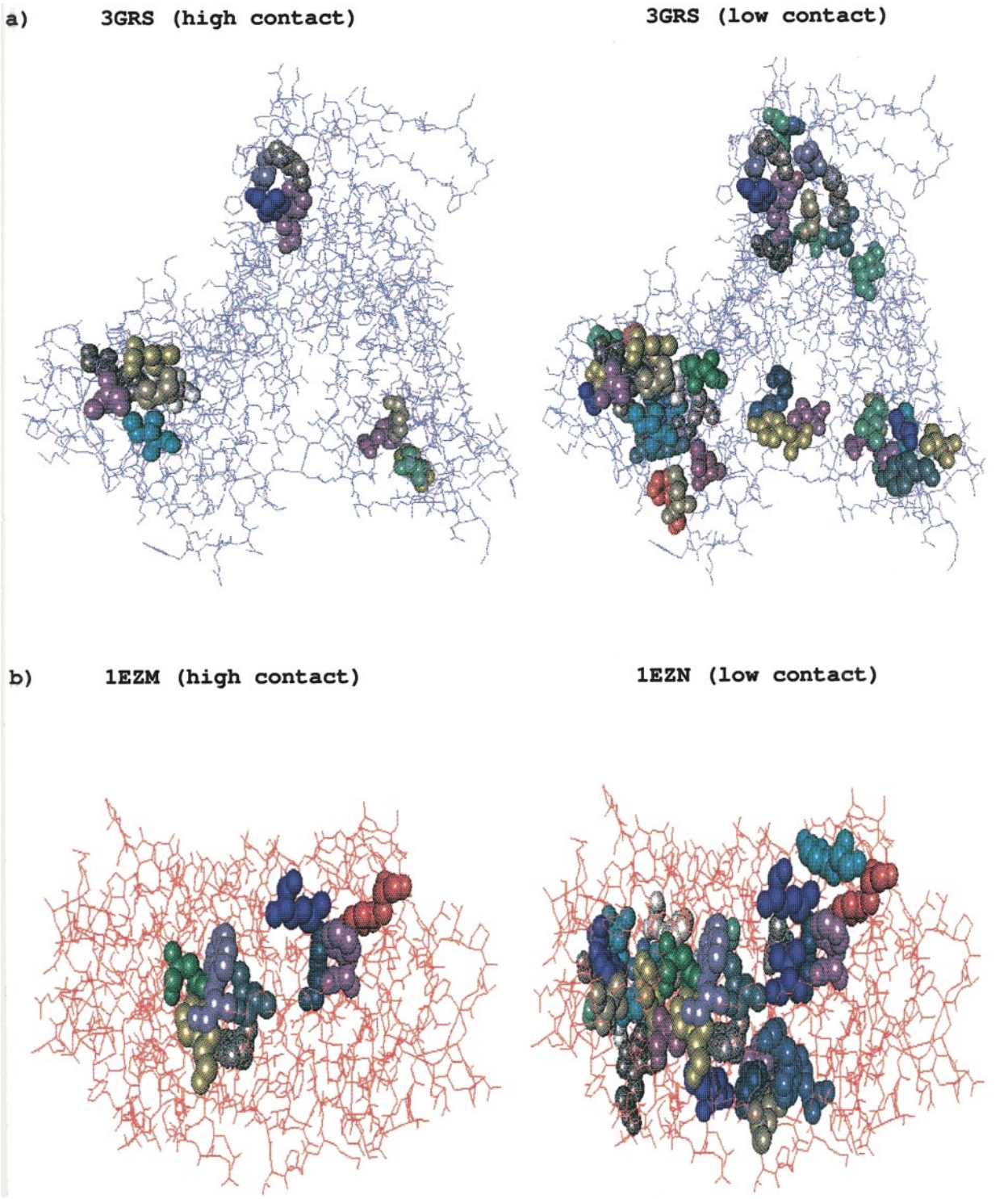

Figure 9. Left panel: The detected cores in glutathione reductase(3GRS) (Karplus & Schulz, 1987) and hydrolase elastase(1EZM) (Thayer *et al*., 1991) using high contact hydrophobic criteria. Right panel: Expanded clusters using low contact hydrophobic criteria. The clusters are shown in VDW representation and the protein molecule in LINE representation.

problems like identification of surface hydrophobic patches involved in protein-protein interaction, specific side-chain networks involved near the active site and glycosylation sites can be probed by this approach and we have started our preliminary investigations on these lines.

Conclusions

A novel technique based on the graph spectral method has been developed for cluster analysis in protein structures. The non-bonded side-chain interaction between residues in proteins and the degree of connectivity of residues are coded in the

form of a matrix called the Laplacian matrix. Side-chain clusters and the information regarding the branching of individual side-chains forming the cluster is obtained from the eigenvalues and eigenvectors of the Laplacian matrix. A variety of clusters are detected by using different side-chain interaction criteria. The developed computer program is flexible enough to obtain different user defined criteria as input and detect clusters of interest. The output of the program is a simple two-dimensional cluster plot and has information on the number of clusters, the type of residues and the accessible surface area of the side-chain forming the cluster. The residues identified as the center of the cluster from the higher eigenvector components are also shown on the cluster plot.

In all the proteins studied, a cluster close the active or binding site was detected, of which at least one residue was in direct interaction with the ligand. This cluster should be important from the structure and function point of view as the interactions of the side-chains forming the cluster should be important to orient the side-chain which is directly involved in the substrate binding. The most interesting feature of the clusters close to the active site is the way in which they expand as the side-chain overlap criteria is reduced. Such expanded clusters are also significant in the context of nucleation-condensation hypothesis of protein folding.

The detected hydrophobic clusters show a good correlation with the experimentally observed folding intermediates. The magnitude of the vector components corresponding to the higher eigenvalues correlate with the Φ values which is a direct measure of the importance of specific side-chain interactions in the transition state of folding. Thus, the method serves the purpose of detecting side-chain clusters which could have formed early during the process of folding and can serve as a tool for studying protein folding.

The algorithm is also useful in detecting domains in protein structures and conserved hydrophobic side-chain clusters in topologically equivalent protein structures. This method can serve as a tool in the analysis of new structures and has potential in detecting surface clusters involved in protein-protein and protein-substrate interaction.

Acknowledgements

We are thankful to the following people from the Indian Institute of Science, Bangalore: Professor Nagendra of Electrical Engineering department and V.S. Anil Kumar of Computer Science and Automation department for useful discussions on graph spectra and graph algorithms, Dr S.M. Patra of Molecular Biophysics Unit for useful suggestions during the course of the work and Professor N.V. Joshi of Center for Ecological Sciences for valuable comments on the manuscript. The computational facilities provided by the Super Computer Education and Research Center and Distributed Information Center of Indian Institute of Science are acknowledged.

References

Anderson, J. E., Ptashne, M. & Harrison, S. C. (1987). Structure of the repressor-operator complex of bacteriophage 434. *Nature,* **326**, 846-852.

Artymiuk, P. J., Rice, D. W., Mitchell, E. M. & Wilett, P. (1990). Structural resemblance between the families of bacterial signal transduction proteins and of G proteins revealed by graph theoretical techniques. *Protein Eng.* **4**, 39-43.

Artymiuk, P. J., Grindley, H. M., Park, J. E., Rice, D. W. & Willett, P. (1992). Three dimensional structural resemblance between leucine aminopeptidase and carboxypeptidase A revealed by graph theoretical techniques. *FEBS Letters,* **303**, 48-52.

Artymiuk, P. J., Poirrette, A. R., Grindley, H. M., Rice, D. W. & Willet, P. (1994). A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains on protein structure. *J. Mol. Biol.* **243**, 327-344.

Bashford, D. W., Chothia, C. & Lesk, A. M. (1987). Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.* **196**, 199-216.

Baudet, S. & Janin, J. (1991). Crystal structure of a barnase-d(GpC) complex at 1.9 Å resolution. *J. Mol. Biol.* **219**, 123-132.

Beardsley, D. S. & Kauzmann, W. (1996). Local densities orthogonal to beta sheet amide planes: Patterns of packing in globular proteins. *Proc. Natl Acad. Sci. USA,* **93**, 4448-4453.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank. A computer-based archival file for macromolecular structures. *Eur. J. Biochem.* **82**, 319-324.

Boppana, R. B. (1987). Eigenvalues and graph bisection: an average-case analysis. *Proc. IEEE Symp. Found. Computer Sci.* 280-285.

Buck, M., Redford, S. E. & Dobson, C. M. (1993). Partially folded state of hen egg white lysozyme in trifluroethanol: structural characterization and implications for protein folding. *Biochemistry,* **32**, 669-678.

Chen, L., Durley, R. C. E., Mathews, F. S. & Davidson, V. L. (1994). Structure of an electron transfer complex: methylamine dehydrogenase, amicyanin and cytochrome c551i. *Science,* **264**, 86-89.

Chou, K. C., Nemethy, G. & Scheraga, H. A. (1990). Energetics of interactions of regular structural elements in proteins. *Accts Chem. Res.* **23**, 134-141.

Chyan, C. L., Wormald, C., Dobson, C. M., Evans, P. A. & Baum, J. (1993). Structure and stability of the molten globule state of guinea-pig alpha-lactalbumin: a hydrogen exchange study. *Biochemistry,* **32**, 5651-5691.

Connolly, M. (1993). The molecular surface package. *J. Mol. Graph.* **11**, 139-141.

Creighton, T. E. & Chothia, C. (1989). Selecting buried residues. *Nature,* **339**, 14-15.

Cvetkovic, D. M. & Gutman, I. (1977). Note on branching. *Croat. Chem. Acta,* **49**, 105-121.

Engelhard, M. & Evans, P. A. (1996). Experimental investigation of side-chain interactions in early folding intermediates. *Folding Design,* **1**, R31-R37.

Evans, P. A., Topping, K. D., Woolfson, D. N. & Dobson, C. M. (1991). Hydrophobic clustering in non native state of a protein interpretation of chemical shifts in NMR spectra of denatured states

of lysozyme. *Proteins: Struct. Funct. Genet.* **9**, 248-266.
Fersht, A. R. (1984). Basis of biological specificity. *Trends Biochem. Sci.* **9**, 145-147.
Fersht, A. R. (1987). Dissection of structure and activity of the tyrosyl-tRNA synthetase by site-directed mutagenesis. *Biochemistry,* **26**, 8031-8037.
Fersht, A. R. (1993). The sixth Datta Lecture. Protein folding and stability: the pathway of folding of barnase. *FEBS Letters,* **325**, 5-16.
Fersht, A. R. (1997). Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3-9.
Garbers, J., Promel, H. J. & Steger, A. (1990). Finding clusters in VLSI circuits. *Proc. IEEE Int. Conf. on Computer-Aided Design,* 520-523.
Gronenborn, A. M. & Clore, G. M. (1994). Experimental support for the ''hydrophobic zipper'' hypothesis. *Science,* **263**, 536.
Guss, J. M. & Freeman, H. C. (1983). Structure of oxidized polar plastocyanin at 1.6 Å resolution. *J. Mol. Biol.* **169**, 521.
Hagen, L. & Kahng, A. B. (1992). New spectral methods for ratio cut partitioning and clustering. *IEEE Trans. Comp. Design,* **11**, 1074-1084.
Hall, K. M. (1970). An r-dimensional quadratic placement algorithm. *Manag. Sci.* **17**, 219-229.
Harbury, P. B., Zhang, T. & Kim, P. S. (1993). A switch between two three and four stranded coiled coils in GCN4 leucine zipper mutants. *Science,* **262**, 1401-1407.
Herinaga, J. & Argos, P. (1991). Side-chain clusters in protein structures and their role in protein folding. *J. Mol. Biol.* **220**, 151-171.
Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522-524.
Holm, L. & Sander, C. (1994). The FSSP database of structurally aligned protein fold families. *Nucl. Acids Res.* **22**, 3600-3609.
Honig, B. & Yang, A. S. (1995). Free energy balance in protein folding. *Advan. Protein Chem.* **46**, 27-59.
Humphrey, W., Dalke, A. & Schulten, K. (1996). VMD - Visual molecular dynamics. *J. Mol. Graph.* **14.1**, 33-38.
Islam, S. A., Luo, J. & Sternberg, M. J. E. (1995). Identification and analysis of domains in proteins. *Protein Eng.* **8**, 513-525.
Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260-288.
Jennings, P. A. & Wright, P. E. (1993). Formation of a molten globule intermediate early in the kinetic pathway of apomyoglobin. *Science,* **262**, 892-896.
Jones, D. H., McMillan, A. J. & Fersht, A. R. (1985). Reversible dissociation of dimeric tyrosil-tRNA synthetase by mutagenesis at the subunit interface. *Biochemistry,* **24**, 5852-5857.
Jones, S., Stewart, M., Michie, A., Swindells, M. B., Orengo, C. & Thornton, J. M. (1998). Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci.* **7**, 233-242.
Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure-pattern recognition of hydrogen bonded and geometrical features. *Biopolymers,* **22**, 2577-2637.
Karlin, S. & Zhu, Z. Y. (1996). Characterizations of diverse residue clusters in protein three-dimensional structures. *Proc. Natl Acad. Sci. USA,* **93**, 8344-8349.
Karplus, P. A. & Schulz, G. E. (1987). Refined structure of glutathione reductase at 1.54 Å resolution. *J. Mol. Biol.* **195**, 701-729.
Koch, I., Kaden, F. & Selbig, J. (1992). Analysis of sheet topologies by graph theory methods. *Proteins: Struct. Funct. Genet.* **12**, 314-323.
Laurence, T. C. & Evans, P. A. (1995). Conserved structural features on protein surfaces: small exterior hydrophobic clusters. *J. Mol. Biol.* **249**, 251-258.
Levinthal, C. (1969). How proteins fold graciously. In *Mossbauer Spectroscopy in Biological Systems* (Debrunner, P., Tsibris, J. C. M. & Munch, E., eds), pp. 22-24, University of Illinois Press, Urbana.
Lu, J. & Dahlquist, F. W. (1992). Detection and characterization of an early folding intermediate of T4 lysozyme using pulsed hydrogen exchange and two dimensional NMR. *Biochemistry,* **31**, 4749-4756.
Lumb, K. J. & Kim, P. S. (1994). Formation of a hydrophobic cluster in denatured bovine pancreatic trypsin inhibitor. *J. Mol. Biol.* **236**, 412-420.
Matouschek, A., Serrano, L. & Fersht, A. R. (1992). The folding of an enzyme structure of an intermediate in the refolding of barnase analysed by a protein engineering procedure. *J. Mol. Biol.* **224**, 819-835.
Mitchell, E. M., Artymiuk, P. J., Rice, D. W. & Willett, P. (1990). Use of techniques from graph theory to compare secondary structural motifs in proteins. *J. Mol. Biol.* **212**, 151-166.
Mossing, M. C. & Sauer, R. T. (1990). Stable, monomeric variants of lambda-Cro obtained by insertion of a designed beta-hairpin sequence. *Science,* **250**, 1712-1715.
Murzin, A. G., Lesk, A. M. & Chothia, C. (1992). Beta-trefoil fold. Patterns of structure and sequence in the Kuintz inhibitors, interleukins and fibroblast growth factors. *J. Mol. Biol.* **223**, 531-543.
Neira, J. L., Davis, B., Ladurner, A. G., Buckle, A. M., de P., Gay G. & Fersht, A. R. (1996). Towards the complete structural characterization of a protein folding pathway: the structures of the denatured, transition and native states for the association/folding of two complementary fragments of cleaved chymotrypsin inhibitor 2. Direct evidence for a nucleation-condensation mechanism. *Folding Design,* **1**, 189-208.
Nemethy, G. & Scheraga, H. A. (1979). A possible folding pathway of bovine pancreatic RNase. *Proc. Natl Acad. Sci. USA,* **76**, 6050-6054.
Ng, K. K., Drickamer, K. & Weis, W. I. (1996). Structural analysis of monosaccharide recognition by rat liver mannose-binding protein. *J. Biol. Chem.* **271**, 663-674.
Patra, S. M. & Vishveshwara, S. (1998). Classification of polymer structures by a graph theory. *Int. J. Quantum Chem.* **71**, 349-356.
Pelletier, H. & Kraut, J. (1992). Crystal structure of a complex between electron transfer partners, cytochrome c peroxidase and cytochrome c. *Science,* **258**, 1748-1755.
Pen, Y. & Briggs, M. S. (1992). Hydrogen exchange in native and alcohol forms of ubiquitin. *Biochemistry,* **31**, 11405-11412.
Plochocka, D., Zielenkiewiez, P. & Rabezenko, A. (1988). Hydrophobic micro-domains as structural invariants in proteins. *Protein Eng.* **2**, 115-119.
Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins. Use of packing criteria in the

enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775-791.
Radford, S. E., Dobson, C. M. & Evans, P. A. (1992). The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature* **358**, 302-307.
Randic, M. (1975). Unique numbering of atoms and unique codes for molecular graphs. *J. Chem. Inf. Comp. Sci.* **15**, 105-108.
Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.* **82**, 1-14.
Samudrala, R. & Moult, J. (1997). Handling context-sensitivity in protein structures using graph theory: bonafied prediction. *Proteins: Struct. Funct. Genet.* **1**, 43-49.
Samudrala, R. & Moult, J. (1998). A graph-theoretic algorithm for comparative modeling of protein structures. *J. Mol. Biol.* **279**, 287-302.
Shakhnovich, E., Abkevich, V. & Ptitsyn, O. (1996). Conserved residues and the mechanism of folding. *Nature,* **379**, 96-98.
Siddiqui, A. S. & Barton, G. J. (1995). Continuous and discontinuous domains: an algorithm for automatic generation of reliable protein domain definition. *Protein Sci.* **4**, 872-884.
Singh, J. & Thornton, J. M. (1991). *Atlas of Protein Side-chain Interactions*, Oxford University Press, Oxford.
Sowhamini, R., Rufino, S. D. & Blundell, T. L. (1996). A database of globular protein structural domains: Clustering of representative family members into similar folds. *Folding Design,* **1**, 209-220.
Stout, G. H. & Jensen, L. H. (1968). *X-ray Structure Determination: A Practical Guide*, Macmillan Company, London.
Swindells, M. B. (1995). A procedure for the automatic determination of hydrophobic cores in protein structure. *Protein Sci.* **4**, 93-103.
Takano, T. (1977). Structure of myoglobin refined at 2 Å resolution. *J. Mol. Biol.* **110**, 569.
Thayer, M. M., Flaherty, K. M. & McKay, D. B. (1991). Three-dimensional structure of the elastase of Pseudomonas aeruginosa at 1.5 Å resolution. *J. Biol. Chem.* **266**, 2864-2871.
Udgaonkar, J. B. & Baldwin, R. L. (1990). Early folding intermediate of RNase A. *Proc. Natl Acad. Sci. USA,* **87**, 8197-8201.
vam de Kamp, M., Silvestrini, M. C., Brunoir, M., van Beumen, J., Hali, F. C. & Canters, G. W. (1990). Involvement of the hydrophobic patch of azurin in the electron transfer reactions with cytochrome c551 and nitrite reductase. *Eur. J. Biochem.* **194**, 109-118.
Waldburger, C. D., Jonsson, T. & Sauer, R. T. (1996). Barriers to protein folding: Formation of buried polar interactions is a slow step in acquisition of structure. *Proc. Natl Acad. Sci. USA,* **93**, 2629-2634.
Weaver, D. L. (1992). Hydrophobic interaction between globin helices. *Biopolymers,* **32**, 477-490.
Weis, W. I. & Drickamer, K. (1994). A trimeric structure of C-type mannose binding protein. *Structure,* **2**, 1227-1240.
Wright, C. S., Wright, R. A. & Kraut, J. (1969). Structure of subtilisin BPN AT 2.5 Å resolution. *Nature,* **221**, 235-242.
Young, L., Jernigan, B. L. & Covell, D. G. (1994). A role for surface hydrophobicity in protein-protein recognition. *Protein Sci.* **3**, 717-729.
Zehfus, M. H. (1995). Automatic recognition of hydrophobic clusters and their correlation with protein folding units. *Protein Sci.* **4**, 1188-1202.

Appendix

A Clustering by Graph Spectra

A set of n points (vertices) in space can be connected by m edges to represent a graph and such a graph can be written in the form of an adjacency matrix. The adjacency matrix of the graph is an $n \times n$ matrix where the ijth entry is 1 if i and j are connected and 0 if they are not connected. 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 points which minimizes the function given below (equation (A1)) (Hall, 1970). In this section we follow the notation and the formulation as given by Hall (1970). Given n points and $n \times n$ symmetric adjacency matrix $\mathbf{A}_{ij}$ which gives the connection between points i and point 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,

$$Z = 1/2 \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)^2 \mathbf{A}_{ij} \qquad \text{(A1)}$$

where $\mathbf{A}_{ij}$ is the adjacency matrix, then the one-dimensional problem is to find a row vector $\mathbf{X}' = (x_1, x_2, \ldots x_n)$ 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:

$$X'X = 1 \qquad \text{(A2)}$$

The solution to the above framed problem is as follows:

Let a_i and a'_j be the ith row and jth column sum, respectively of the matrix $\mathbf{A}$. Since the adjacency matrix $\mathbf{A}$ is symmetric $a_i = a'_j$. Define a diagonal matrix $\mathbf{D} = (d_{ij})$ as follows:

$$d_{ij} = \begin{cases} 0 & \text{if} \quad i \neq j \\ a_i & \text{if} \quad i = j \end{cases}$$

Now define the following matrix:

$$\mathbf{B} = \mathbf{D} - \mathbf{A} \qquad \text{(A3)}$$

The matrix $\mathbf{B}$ is called the Laplacian matrix.

Now we shall show the expression (A1) can be rewritten in terms of the Laplacian matrix as:

$$Z = X'BX \qquad \text{(A4)}$$

Expanding equation (A1) we get:

$$Z = 1/2 \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i^2 - 2x_i x_j + x_j^2) \mathbf{A}_{ij} \qquad \text{(A5)}$$

$$Z = 1/2\left(\sum_{i=1}^{n} x_i^2 a_i - 2\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j \mathbf{A}_{ij} + \sum_{j=1}^{n} x_j^2 a'_j\right) \tag{A6}$$

$$Z = \sum_{i=1}^{n} x_i^2 a_i - \sum_{j=1}^{n}\sum_{i\neq j}^{n} x_i x_j \mathbf{A}_{ij} \tag{A7}$$

Since $\mathbf{A}_{ij}$ is a symmetric matrix $a_i = a'_j$, hence Z can be written as:

$$Z = X'BX \tag{A8}$$

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

$$L = X'BX - \lambda(X'X - 1) \tag{A9}$$

Taking the first partial derivative of L with respect to the vector $\mathbf{X}$ and setting the result equal to zero yields:

$$2BX - 2\lambda X = 0 \tag{A10}$$

If I is identified as the identity matrix, the above equation can be rewritten as:

$$(B - \lambda I)X = 0 \tag{A11}$$

which yields a nontrivial solution X, if and only if λ is an eigenvalue of the matrix $\mathbf{B}$ and $\mathbf{X}$ is the corresponding eigenvector. If the above equation is premultiplied by X' and the constraint equation (A2) is applied we obtain:

$$\lambda = X'BX \tag{A12}$$

Thus, the formal solution to equations (A1) and (A2) is simply that X is the eigenvector of B 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. As we shall see, this solution is related to the clustering of points. The above solution for one dimension also holds good in two and three-dimensional space (Hall, 1970).

The above given procedure for clustering by graph spectra is illustrated by the following example. Consider a set of points in two-dimensional space as given in Figure A1(a). In order to cluster the points in two-dimensional space, a graph for the given configuration of points is constructed. The points would represent the vertices of the graph and the edges would represent $1/d_{ij}$, where d_{ij} is the distance between the points i and j as given in Figure A1(b). The reciprocal of the distances ensures that higher edge weights are assigned to shorter distances. Here, only those points that are less than a distance of three units ($1/d_{ij} \geqslant 0.33$) are considered to be connected in the graph. An extremely low edge weight of 0.01 is assigned to those points that are separated by a distance greater than three units.

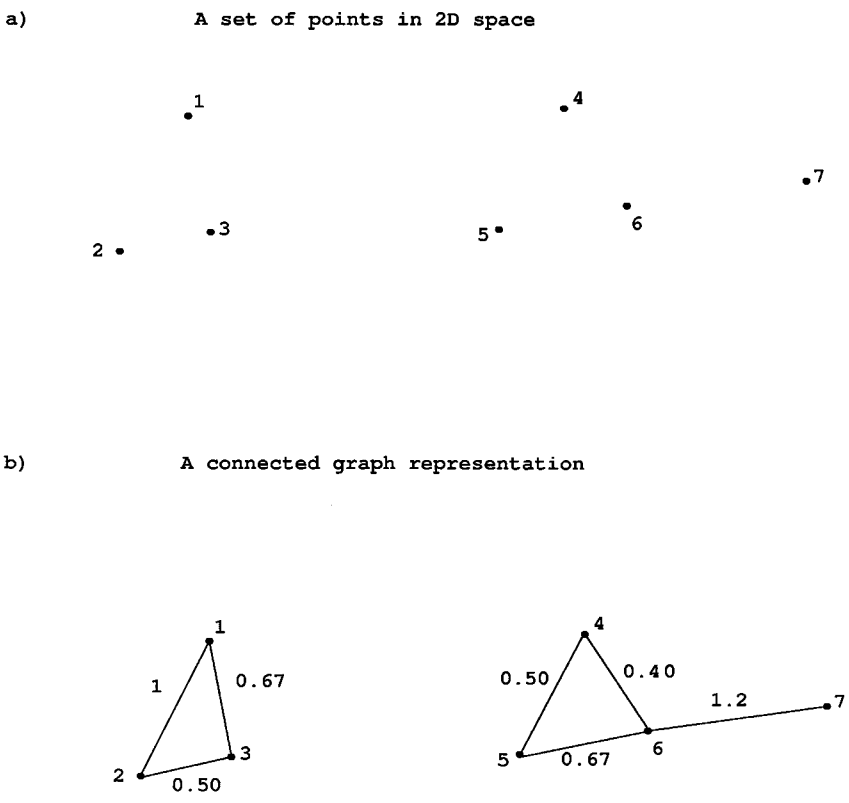

Figure A1. (a) A set of points in two dimensional space. (b) A connected graph representation for the configuration. The vertices are numbered and the edges are shown by their corresponding edge weight.

One can use any edge weight much less than ($\frac{1}{3}$) for any two vertices greater than a distance of three units.

Now, the adjacency matrix **A** for the graph in Figure A1(b) is given by:

$$A = \begin{pmatrix} 0 & 1 & 0.67 & 0.01 & 0.01 & 0.01 & 0.01 \\ 1 & 0 & 0.50 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.67 & 0.50 & 0 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0 & 0.5 & 0.4 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.5 & 0 & 0.67 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.4 & 0.67 & 0 & 1.2 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 1.2 & 0 \end{pmatrix}$$

and the degree matrix **D** is given by:

$$D = \begin{pmatrix} 1.71 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.54 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.21 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.94 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.21 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2.3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1.25 \end{pmatrix}$$

Hence the Laplacian is given by $\mathbf{B} = \mathbf{D} - \mathbf{A}$ which is:

$$B = \begin{pmatrix} 1.71 & -1 & -0.67 & -0.01 & -0.01 & -0.01 & -0.01 \\ -1 & 1.54 & -0.50 & -0.01 & -0.01 & -0.01 & -0.01 \\ -0.67 & -0.50 & 1.21 & -0.01 & -0.01 & -0.01 & -0.01 \\ -0.01 & -0.01 & -0.01 & 0.94 & -0.5 & -0.4 & -0.01 \\ -0.01 & -0.01 & -0.01 & -0.5 & 1.21 & -0.67 & -0.01 \\ -0.01 & -0.01 & -0.01 & -0.4 & -0.67 & 2.3 & -1.2 \\ -0.01 & -0.01 & -0.01 & -0.01 & -0.01 & -1.2 & 1.25 \end{pmatrix}$$

The Laplacian matrix is diagonalized and the eigenvalues and eigenvectors are tabulated below.

Eigenvalues	Vertex no.	0.0000	**0.0700**	0.8342	1.5786	1.7697	2.6503	3.2573
Eigenvectors	1	0.3780	**0.4364**	−0.0000	0.0000	0.2824	**−0.7661**	−0.0000
	2	0.3780	**0.4364**	−0.0000	0.0000	0.5223	0.6276	0.0000
	3	0.3780	**0.4364**	−0.0000	0.0000	−0.8047	0.1385	−0.0000
	4	0.3780	**−0.3273**	−0.5959	0.6223	−0.0000	0.0000	0.0876
	5	0.3780	**−0.3273**	−0.3364	−0.7587	0.0000	−0.0000	0.2475
	6	0.3780	**−0.3273**	0.2457	−0.0498	0.0000	0.0000	**−0.8289**
	7	0.3780	**−0.3273**	0.6866	0.1862	−0.0000	−0.0000	0.4939

It is evident that the lowest eigenvalue 0.000 gives rise to a redundant solution where all the components have a value of 0.378 which is $1/\sqrt{7}$ since there are seven vertices (points). An examination of the components of the eigenvector corresponding to the second lowest eigenvalue shows that they form two distinct graphs (clusters): 1, 2 and 3 have the same value (0.4364) and components 4, 5, 6 and 7 have the same value (−0.3273). The second lowest eigenvalue (0.0700) and its corresponding vector components are shown in bold. It is also evident that the vector components of the top eigenvalues (3.257 and 2.650) have information on any one of the two clusters as the vector components are non-zero for the vertices which form a cluster. From Figure 1(b) it is observed that vertex 6 is highly branched as compared to other vertices which form cluster 2. This information is derived directly by examining the vector components of the top eigenvalue (3.257) of which vertex 6 has the largest magnitude of 0.8289. In the case of cluster 1 formed by vertices 1, 2 and 3 all the vertices have equal degree of two, but vertex 1 has the two largest edge weights (1 and 0.67) connected to it (Figure A1(b)). This information is directly obtained from the components of the second largest eigen value (2.650) of which the magnitude of vertex 1 is high (0.7661) as compared to the other two vertices 2 and 3.

Thus the above simple example demonstrates how the analysis of eigenvalues and eigenvectors called the study of graph spectra can be used for deriving information regarding clusters and cluster centers

References

Hall, K. M. (1970). An r-dimensional quadratic placement algorithm. *Manag. Sci.* **17**, 219-229.

Edited by J. M. Thornton

(Received 1 March 1999; received in revised form 12 July 1999; accepted 20 July 1999)