ELSEVIER

Biophysical Chemistry 84 (2000) 13–25

Biophysical Chemistry

www.elsevier.nl/locate/bpc

Backbone cluster identification in proteins by a graph theoretical method

S.M. Patra, S. Vishveshwara*

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

Received 19 July 1999; received in revised form 9 October 1999; accepted 2 November 1999

Abstract

A graph theoretical algorithm has been developed to identify backbone clusters of residues in proteins. The identified clusters show protein sites with the highest degree of interactions. An adjacency matrix is constructed from the non-bonded connectivity information in proteins. The diagonalization of such a matrix yields eigenvalues and eigenvectors, which contain the information on clusters. In graph theory, distinct clusters can be obtained from the second lowest eigenvector components of the matrix. However, in an interconnected graph, all the points appear as one single cluster. We have developed a method of identifying highly interacting centers (clusters) in proteins by truncating the vector components of high eigenvalues. This paper presents in detail the method adopted for identifying backbone clusters and the application of the algorithm to families of proteins like RNase-A and globin. The objective of this study was to show the efficiency of the algorithm as well as to detect conserved or similar backbone packing regions in a particular protein family. Three clusters in topologically similar regions in the case of the RNase-A family and three clusters around the porphyrin ring in the globin family were observed. The predicted clusters are consistent with the features of the family of proteins such as the topology and packing density. The method can be applied to problems such as identification of domains and recognition of structural similarities in proteins.

Keywords: Non-covalent interaction; Adjacency matrix; Eigenvalue and eigenvector; Cluster; RNase-A family; Globin family; Domains

* Corresponding author. Tel.: +91-80-3092611; fax: +91-80-3348535.
E-mail address: sv@mbu.iisc.ernet.in (S. Vishveshwara)

0301-4622/00/$ - see front matter
PII: S0301-4622(99)00134-9

1. Introduction

The problem of protein folding has been a subject of great interest for many theoreticians and experimentalists [1,2] for the last two decades. The key issue is to unravel the code for the unique three-dimensional structure of proteins. One of the methodologies adopted is to analyze the crystal structures of proteins and derive knowledge-based rules [3].

It is clear from many observations that non-local interactions play a major role in the process of protein folding and in deciding the three-dimensional structures of proteins [4,5]. Protein structures are dominated by non-covalent interactions with non-uniform packing density [6]. Clustering of amino acid residues are seen in proteins [3,7]. Analysis of clusters gives insight into the folding patterns and possibly to the folding pathways in proteins. The clusters may be formed at some stage in the folding process, which considerably reduces the searchable conformational space and drives protein folding in the correct direction. The clusters also provide hints for site-directed mutagenesis and protein engineering experiments as they are important for structural stability. The ultimate goal in protein structure prediction is to deduce the three-dimensional structure of any protein purely from its sequence. The interactions among residues, together with their interactions with the surrounding medium determine the unique structure of globular proteins. Therefore, any cluster of interacting residues could be of significant interest from the viewpoint of protein folding and stability. Several cluster identification algorithms are available in the literature [3,7–10]. Each of the methods highlight different aspects of protein structure.

Graph theory in general is an important tool in topological investigation. Proteins have been investigated earlier by graph theoretical methods in contexts such as secondary structure identification [11], recognition of β-folds in β-barrels [12] and tertiary folds in G-proteins [13] from sequence information. Bahar et al. [14–17] have applied Kirchoff's matrix to describe the spatial neighboring residues in proteins and have elucidated a number of properties such as vibrational dynamics and thermal fluctuations in proteins. In this paper we applied this technique on the three-dimensional structure of proteins to identify non-bonded clusters from the contacts between spatially neighboring amino acid residues in proteins. The advantage of this method is that the clusters are automatically obtained as solutions to the connectivity matrix. Computationally, it is an elegant procedure and various features such as the connectivity between clusters, the basic core (core corresponds to the minimum number of residues present in each clusters) of the cluster and the core expansion can be easily obtained from the solution. Earlier, for the first time we adopted this procedure to identify non-bonded clusters in two- and three-dimensional lattice models [18]. The present application to real protein structures enables us to detect conformational similarities by relating eigenvalues and vectors to a cluster of connected interactions. Recently, an elegant method has been described to identify side chain clusters in proteins [19]. In the present study, the backbone C_α atom is used to define the connectivity and therefore, the resulting clusters are referred to as backbone clusters. In general, the core obtained in the present study is similar to the stabilization center (SC) described in earlier studies [7].

2. Method

2.1. Construction of adjacency matrix

Graphs can be represented in algebraic form as matrices [20]. The molecular graph is a non-numerical representation of the chemical structure. The presentation of graphs in matrices allows the computerised manipulation of graphs through the agency of standard numerical algorithms. The adjacency matrix $A = A\ (G)$ of a graph G with N vertices is the square N^*N symmetric matrix.

$[A]_{ij} = 1$ if $i \neq j$ and $e_{ij} \varepsilon\ E\ (G)$

0 if $i = j$ or $e_{ij}\ \varepsilon\ E\ (G)$

[Where e_{ij} is edge between i and j and $E(G)$ is edge set of graph G]

The first step of our method is to construct an adjacency matrix of the backbone structure of a protein whose co-ordinates are obtained from the PDB file. The matrix is constructed by identifying the spatial neighbours lying within the range of 6.5 Å from the selected atom in the protein [21]. The number of C_α atoms surrounding the chosen C_α (C_i) along the chain (excluding the two sequence neighbors C_{i-2}, C_{i-1}, C_{i+1} and C_{i+2}) are identified as spatial neighbors for each residue of the protein and this number is designated as contacts. We have chosen a sphere of radius 6.5 Å, which was found to be approximately the distance corresponding to the first peak in the radial distribution of residues in the interior of proteins [22,23]. Panjikar et al. [21], have used this concept in determining backbone packing in globular proteins. We have constructed an adjacency matrix in which the C_α atoms of residues i and j are within 6.5 Å. The matrix element $a_{ij} = 1$ if $C_{\alpha i}$ and $C_{\alpha j}$ are within 6.5 Å. All a_{ij} with distance ($C_{\alpha i}$–$C_{\alpha j}$) > 6.5 Å are set at zero. Furthermore, the element a_{ik} is also set at zero, where $k = i - 2$ to $i + 2$.

2.2. Diagonalization of matrix and sorting of eigenvalue and eigenvector components

The general matrix diagonalization equation is given by $(A - \lambda I)\ X = 0$, where I is the identity matrix, X is the eigenvector and A is the adjacency matrix. The matrix A is diagonalized using MATLAB (version 5.0, 1996) to obtain eigenvalues and eigenvectors. The eigenvalues are sorted along with their corresponding eigenvector components [24]. The sorted eigenvalues and eigenvectors of such a matrix contain information regarding clusters (highly connected region), the important residues which make the cluster (core) and the nature of connectivity of the residue (branching parameter). We can also obtain the number of edges connected to a center (degree) by eigenvalue analysis. Generally, cluster information can be obtained from the second lowest eigenvector component of the Laplacian matrix [24]. However, if all the points are interconnected, the distinct clusters cannot be identified by the second lowest eigenvector. We have devised a method of identifying distinct clusters in a connected graph using top (higher) eigenvalues. As Gutman [25] has pointed out, the highest eigenvalue is related to the branching parameter. Many physical and chemical properties of saturated hydrocarbons depend on the extent of branching of the carbon skeleton of the molecule. The boundary limit of the highest eigenvalue is $D_{\min} \le \lambda_1 \le D_{\max}$ [25], where $D_{\max}$ and $D_{\min}$ are the maximum and minimum degrees of the graph, respectively and λ_1 is the highest eigenvalue. The present analysis on protein structures is to elucidate the nature of non-covalent interactions, where the maximum branching at an edge ranges from four to nine [21] representing the backbone packing density. The high branching centers correspond approximately to the stabilization center (SC) [7] in proteins. In the graph theoretical context, the eigenvector components of high eigenvalues highlight such stabilization centers. In Section 2.3 we discuss our method for identifying backbone clusters.

2.3. Identification of independent clusters

In the two-dimensional lattice model, the different eigenvalues corresponded to different connected subgraphs, i.e. small clusters [18]. However, in three-dimensional models [18] and in proteins most of the residues are interconnected and hence a clear identification of the cluster by the inspection of eigenvalue is not possible. The eigenvalues and their eigenvector components give an indication of the corresponding individual cluster. Randic [26] has shown that the largest coefficient belongs usually to the vertices of maximal degree, while the smallest coefficient belongs to terminal vertices and their neighbors and this was found to be true in the non-covalent interactions in polymers in two-dimensional models [18]. However, in proteins the contribution from other clusters is non-zero and hence the problem of terminating a cluster arises. This problem of identifying residues belonging to individual clusters is handled systematically in the following way. In

our observation the highest branching center or the center connected to highly branched centers contributes the highest vector component (X_i) to the corresponding eigenvector of higher eigenvalues and the component values tend towards zero as we move away from the branching center. The eigenvectors are normalized, i.e.

$\sum_{i=1}^{N} (X_i)^2 = 1$, where N is the

total number of centers.

We define another term called sum $(n) = \sum_{i=1}^{n} (X_i)^2$,

Where n = number of sorted eigenvector components, which corresponds to 'n' number of residues and $< N$. A plot of the number of residues vs. sum (n) in the RNase-A (7 rsa) is given in Fig. 1. From the sorted eigenvector components, initially there is a rapid increase of sum (n) and the contribution slows down as we come down the components. In order to select the components contributing to the rapid increase of sum (n) we have adopted a simple method, in which we start by taking at a minimum the two highest eigenvector components (designated as cutpoint 2) and a regression line is drawn. The regression line where it cuts the X axis gives the total number of residues included in the cluster at cutpoint 2 and we define the number of residues present in a cluster at a minimum cutpoint as the core of the cluster. As the cutpoint increases the size of the cluster increases, including more and more numbers of residues.

The vectors corresponding to high eigenvalues may represent independent stabilization centers or may be a different representation of the same center. Our goal is to pick up all independent centers in the protein. We select approximately 10 top eigenvectors and their corresponding components. The rationale behind this selection is based on our observation on the proteins investigated, that the top eigenvalues and their vector components contain the information relevant to independent clusters. The lower eigenvalues do not offer any information on new clusters. Approximately 10 eigenvectors of the sorted high

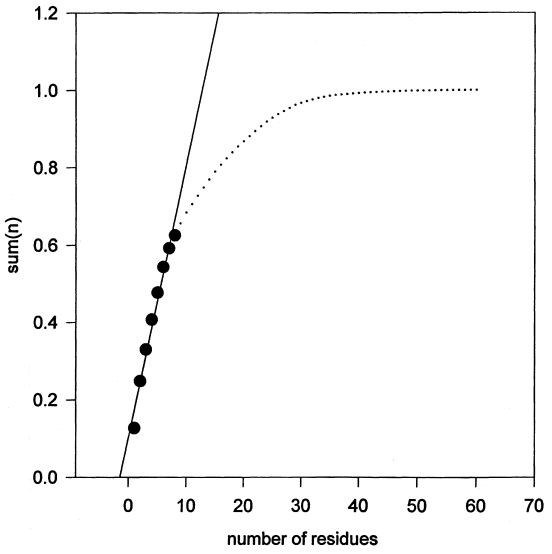

Fig. 1. A plot of sum (n) [see Section 2 for details] (Y axis) vs. the number of residues (X axis) corresponding to the highest eigenvector of 7 rsa. A regression line (—) of eight points (●) intersects the X axis at 13 (number of residues).

eigenvalues are selected and designated as top eigenvectors. In each one of them we identify the stabilization center by the minimum cut point method as described above. Then a cross-check of the residues, which have been obtained from various eigenvectors is done. If the selected residues figure out in only one eigenvector, it is considered as an independent cluster. If some of the selected residues appear in more than one eigenvector, the eigenvector with the higher eigenvalue is considered.

The cluster identification method discussed here is illustrated with an example of the protein hemoglobin which contains 136 amino acid residues. The co-ordinate selection is done from PDB (1 eco) file and adjacency matrix is constructed by checking the distance between C_α atoms among themselves. The matrix element $A_{(ij)}$ is considered to be 1 if the distance between the $C_{\alpha i}$ and $C_{\alpha j}$ atoms is less than 6.5 Å (excluding the cases where j takes the values between $i-2$ and $i+2$) and all other matrix elements are considered to be zero. The matrix (of the order 136) thus obtained is diagonalized and eigenvalues and eigenvectors are sorted. The top 10

eigenvectors are compared for independent clusters as follows. The number and nature of residues identified by the minimum cutpoint method (core residues) in each of the 10 eigenvectors are given in Table 1. These core residues are compared between themselves in each of the 10 eigenvectors. The number of common residues among 10 eigenvectors is given in Table 2 in a matrix form. In Table 2 the diagonal entries represent the number of residues present in that particular

Table 1
Core residues and eigenvector coefficients of top 10 eigenvectors for 1 eco

Eigenvalue	Corresponding vector component with residue								
4.774	0.246	0.246	0.282	0.288	0.289	0.292	0.336	0.361	
	30 A	100 F	26 A	23 I	104 F	107 Y	27 V	103 G	
4.186	0.182	0.211	0.226	0.288	0.298	0.338	0.343	0.415	
	123 A	124 T	8 T	5 Q	121 W	117 A	120 A	9 V	
4.092	0.149	0.167	0.245	0.247	0.254	0.273	0.307	0.322	0.408
	62 I	103 G	21 V	56 E	24 L	60 N	63 V	20 P	59 A
3.507	0.198	0.241	0.252	0.316	0.381	0.396			
	41 F	38 F	42 A	32 P	28 F	35 M			
3.443	0.143	0.143	0.152	0.153	0.163	0.164	0.169	0.171	0.172
	88 K	130 G	125 L	35 M	126 D	79 D	87 H	81 N	128 F
	0.202	0.208	0.214	0.228	0.250	0.301			
	129 F	76 I	83 F	84 V	132 I	80 V			
3.279	0.141	0.141	0.143	0.144	0.146	0.148	0.150	0.152	0.153
	8 T	10 Q	82 T	9 V	131 M	87 H	133 F	79 D	81 N
	0.162	0.164	0.170	0.170	0.171	0.183	0.184	0.188	0.188
	80 V	123 A	128 F	83 F	84 V	76 I	122 G	125 L	127 T
	0.200	0.204	0.212	0.236					
	129F	130G	126D	69I					
3.124	0.132	0.135	0.142	0.154	0.159	0.173	0.175	0.186	0.194
	23 I	35 M	64 G	112 T	26 A	68 K	65 F	69 I	98 N
	0.200	0.203	0.221	0.231	0.248	0.288			
	101 R	114 F	102 A	108 M	109 K	105 V			
3.039	0.139	0.142	0.147	0.153	0.156	0.159	0.171	0.173	
	112 T	81 N	70 I	64 G	59 A	84 V	26 A	73 L	
	0.173	0.176	0.179	0.203	0.215	0.219	0.220	0.239	0.268
	101 R	114 F	72 E	109 K	65 F	68 K	108 M	105 V	69 I
2.826	0.130	0.134	0.167	0.182	0.205	0.211	0.212	0.223	0.223
	30 A	97 L	111 H	114 F	31 D	96 Q	104 F	23 I	95 D
	0.231	0.232	0.240	0.250	0.399				
	98 N	108 M	107 Y	102 A	99 N				
2.743	0.164	0.168	0.170	0.179	0.192	0.208	0.218	0.280	0.309
	132 I	129 F	3 A	120 A	13 F	17 K	11 A	14 D	7 S
	0.343								
	10 Q								

Table 2
Number of common residues in the core cluster of 1 eco from the top 10 eigenvectors

	1	2	3	4	5	6	7	8	9	10
1	8	0	1	0	0	0	2	1	4	0
2	0	8	0	0	0	3	0	0	0	1
3	1	0	9	0	0	0	0	1	0	0
4	0	0	0	6	1	0	1	0	0	0
5	0	0	0	1	15	12	1	2	0	2
6	0	3	0	0	12	22	1	3	0	2
7	2	0	0	1	1	1	15	11	5	0
8	1	0	1	0	2	3	11	17	2	0
9	4	0	0	0	0	0	5	2	14	0
10	0	1	0	0	2	2	0	0	0	10

cluster and the off diagonal entries indicate the number of residues that are common between the eigenvectors. The 1st eigenvector has common core residues with the 3rd, 7th, 8th and 9th eigenvectors. The 2nd eigenvector has common residues with the 6th and 10th eigenvectors and the 4th overlaps with the 5th and 7th eigenvectors. Thus, the only three independent eigenvectors are the 1st, 2nd and 4th containing eight, eight and six core residues, respectively. They are represented in the box in Table 2.

When the cutpoint is increased, the cores also expand, and at cutpoint 17 two clusters corresponding to eigenvector 1 and 4 of hemoglobin (1 eco) merge (Table 4) and give rise to only two independent clusters corresponding to eigenvector 1 and 2. These two clusters merge at cutpoint 34. It has been observed in all the cases that merging of clusters takes place at the residue which is either the center of another cluster or is present in the expanding regions of a cluster.

In Section 3, we demonstrate the power of this graphs theoretical approach to identify non-bonded backbone clusters in proteins by identifying the clusters in two families of proteins, RNase-A and globins. The list of selected proteins is given in Table 3. The eigenvectors of the adjacency matrices are generated and sorted as mentioned in Section 2.2. The details of the clusters are presented in Table 4 and the graphical representations are shown in Fig. 2.

3. Results and discussion

3.1. RNase-A family

RNase-A family proteins from different sources have been selected for the present investigation. They carry out different functions, although all of them perform the common ribonucleolytic action [27]. Structurally they are similar. Backbone clusters are analyzed in the present study. The results presented in Table 4 show that the number of clusters vary from two to four. A close inspection of the clusters graphically presented in Fig. 2 shows that there are generally three regions where clusters are formed, one on the left lobe (I), one in the middle (III) and one on the right lobe (II) of the molecule. The middle one is not strong enough to be recognized in 1 agi and the left one is divided into two in 1 onc. A sequence alignment of the selected proteins (using the program,

Table 3
Proteins from the RNase-A and globin families selected for cluster analysis

PDB code	Name	Source	Resolution (Å)
7 rsa	Ribonuclease A	Bovine pancreas	1.26
1 ang	Angiogenin	Human	2.4
1 agi	Angiogenin	Bovine (milk)	1.5
1 onc	Pancreatic ribonuclease	Rana Pipiens	1.7
1 eco	Hemoglobin	Chironomous thumi thumi	2.0
1 mba	Myoglobin	Sea hare	2.0
1 mbs	Myoglobin	Common seal	2.5
1 nih	Hemoglobin	Human	2.6

Table 4
The identified clusters in selected proteins

PDB code	Total no. of residue	Cutpoint number	No. of cluster	No. of residues in each cluster and their corresponding eigenvector in ()
7 rsa	124	2	3	8 (1st) 8 (2nd) 10 (3rd)
		8	2	13, 14
		39	1	39
1 ang	123	2	3	8 (1st) 10 (2nd) 10 (3rd)
		10	2	13, 15
		30	1	31
1 agi	125	2	2	13 (1st) 9 (2nd)
		28	1	31
1 onc	104	2	4	8 (1st) 8 (2nd) 7 (3rd) 8 (4th)
		9	3	13, 12, 12
		20	2	21, 20
		34	1	34
1 eco	136	2	4	8 (1st) 8 (2nd) 6 (4th)
		16	3	18, 18, 21
		17	2	19, 19
		34	1	34
1 mba	146	2	2	13 (1st) 9 (3rd)
		20	1	26
1 mbs	153	2	2	6 (1st) 12 (2nd)
		32	1	32
1 nih	146	2	3	8 (1st) 9 (2nd) 9 (3rd)
		4	2	9, 10
		46	1	46

Multiple Alignment) is given in Table 5 with the cluster residues highlighted. Several interesting features can be noticed from Fig. 2 and Table 5. Cluster I is generally made up of hydrophobic residues such as (P, A, V), whereas a number of polar and charged residues (Y, E/D, K/R, S/T), appear in cluster II. The center (high branching center with highest vector component) of cluster II, is a tyrosine, 97Y (all numberings in this discussion will correspond to that of 7 rsa) in all the proteins. Its importance in folding is shown by mutation studies [28]. The center of cluster I, however, varies. In fact in 1 onc, cluster I region is made up of two core clusters. The active site residues 11Q and 12H are a part of cluster III. The important conserved hydrophobic residues of cluster III precede the binding site residue 45T and similarly the catalytic residue 119H is present immediately after the conserved PV (117–118) residues which is a part of cluster I. Proline 113 in 1 agi (equivalent to P117 of 7 rsa) is in fact the center of cluster I and the equivalent residues in other proteins have high eigenvector components. The importance of disulfide bridges can also be seen in the identified clusters. The disulfide bridges between 58 and 100 C and 26 and 82 C are the important components of cluster I and II, respectively, in all the proteins. On the other hand, 61 C of the disulfide bridge (61–70 C) is not in the core cluster and perhaps is not essential to

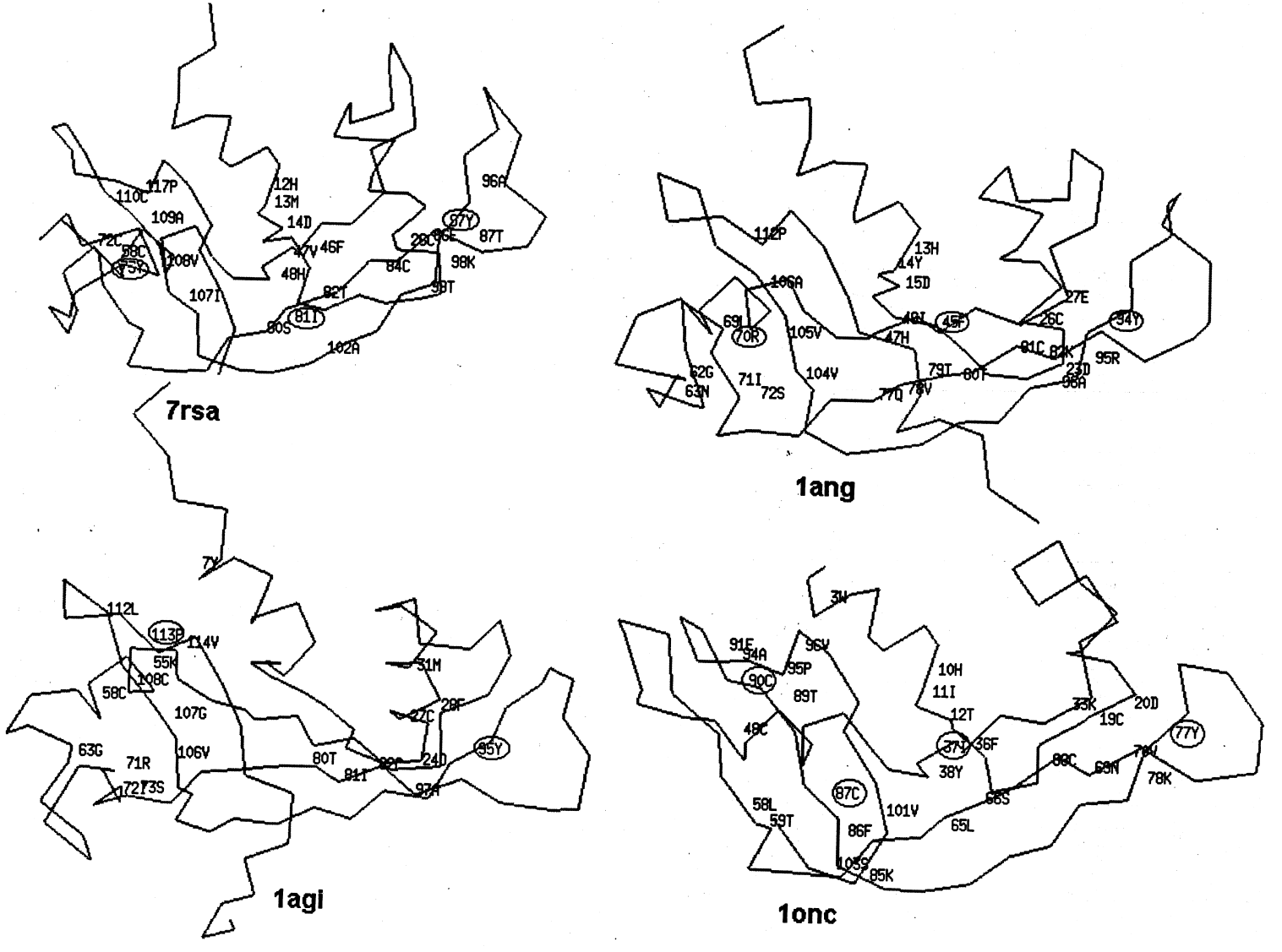

Fig. 2. A graphical representation (generated by VMD [30]) of different core clusters with their centers (highest vector component marked by O) for RNase-A family proteins.

Table 5
Sequence alignment of RNase-A family

The highlighted residues of RNase-A family appear in the core clusters as shown in Fig. 2							
1 ang	.QDNSRYTHF	LTQ**HYD**AKPQ	.GR**D**DRY**CE**S	IMRRRGLTSP	.CKDINT**FIH**	GNKRSIKAIC	ENKN**G**NPHRE
1 agi	AQDDYR**Y**IHF	LTQHYDAKPK	.GRN**D**EY**C**FN	M**M**KNRRLTRP	.CKDRNTFIH	GNKND**I**KAI**C**	EDRN**G**QPYRG
7 rsa	~ ~ KETAAAKF	ERQ**HMD**SSTS	AASSSNY**C**NQ	MMKSRNLTKD	RCKPVNT**FVH**	ESLADVQAV**C**	SQKNVACKNG
1 onc	~ ~ ~ ~ ZD**W**LTF	QKK**HIT.....**	.NTRDVD**CD**N	I**MS**TNLF...	HCKDKNT**FIY**	SRPEPVKAI**C**	KGIIAS....
1 ang	..N**LRI..S**KSS	F**QVTTCK**LHG	GSPWPPCQ**YR**	**A**TAGFRNV**VV**	**A**CENG..L**PV**	H**L**DQSIFRRP	~
1 agi	..DL**RI..S**KSE	FQI**TIC**KHKG	GSSRPPCR**Y**G	**A**TEDSRVIV**V**	**GC**ENG..**LPV**	HFDESFITPR	H
7 rsa	QTN**CY**Q.SYST	M**SIT**D**C**R**ET**G	SSKYPNC**AYK**	**T**TQ**A**NKHI**IV**	**AC**EGNPYV**P**V	HFDASV ~ ~ ~ ~	~
1 onc	..KNVLTTSE	FY**LS**D**CNVT**S	R....PC**KYK**	LKKSTN**KF**CV	**TC**EN..Q**APV**	HFVG**V**G**SC** ~ ~	~

hold the structure. This idea is also supported by its absence in angiogenin and complete lack of this loop structure in 1 onc. The role of the fourth disulfide bridge (40–95 C) seems to be to hold the two loops in position and to place 97Y in proper place for cluster formation. This disulfide by itself does not become part of the core cluster. Thus, the signature of RNase-A family proteins is also seen at the core cluster level with many of the conserved and active site residues taking part in the core cluster formation.

3.2. Globin family

Globins are α-helical proteins consisting of helices (A–H). The prosthetic group porphyrin is surrounded by three regions, which we have marked as I, II and III in Fig. 3. The residues participating in these clusters are also presented in Fig. 3 and the number of independent clusters are listed in Table 3. Several helices cross each other [29] in different regions of globins, for instance the crossing of helices B–E, B–G occurs in the region marked as I and the crossing of helices H–F–G and A–H takes place in regions II and III, respectively. As noted earlier [21] these helix crossings occur at places where small residues like glycine, alanine and serine are present with high backbone neighbor density (denoted as contacts [21]). Interestingly, these are the regions which show up as core clusters in our present analysis. These clusters for some of the globins are presented in Fig. 3. The residues constituting the top three eigenvector components in these clusters are presented in Table 6. They invariably appear in high neighbor density regions as seen by high contact number. In most of the cases, the clusters are made up of two helices and in some cases from three helices (1 mba, II cluster). Apart from the small residues, glycine, alanine and serine, occasionally valine and methionine also appear in these core clusters and these residues have the highest eigenvector component of the corresponding cluster. While the first highest vector component appears from one helix, the second or third highest component comes from the other crossing helix. This indicates a high branching factor for the residues of both the helices in the crossing region. The high vector component residues in a given cluster is similar to the stabilization center (SC) identified earlier [7]. It was earlier stated that [7] the SCs tend to connect more sheets to each other, while connections with helices and turns are less populated. However, helical proteins like globins contain only helices as secondary structures which need to be stabilized. In the present study we have found that the interhelical crossing regions correspond to such SC in globins.

In the present study we have thus demonstrated that backbone clusters in proteins can be identified by graph theoretical parameters and that the clusters in similar regions of structurally similar proteins can be detected. The method, however, can be applied to other interesting problems such as the identification of domains and

Table 6
Core clusters of globin family (Fig. 3): residues corresponding to the three highest vector components from the corresponding eigenvector

CLUSTER	IA	IB	II	III
1 eco	35M (5)[a], 28F (6) 32P (4)	103G (7) 27V (6), 107Y (6)		94V (7) 120A (5) 117A (5)
1 mba	64S (7), 25G (6) 26L (6)		141L (8), 92A (7) 102F (6)	
1 mbs	25G (10), 65G (9) 26Q (8)			131M (6) 10V (7), 132K (6)
1 nih	24G (7), 64G (7) 65K (8)	109V (6) 27A (7), 113V (5)	89S (8), 85S (6), 140A (6)	

[a] The values in parentheses are the number of contacts made by the residue.

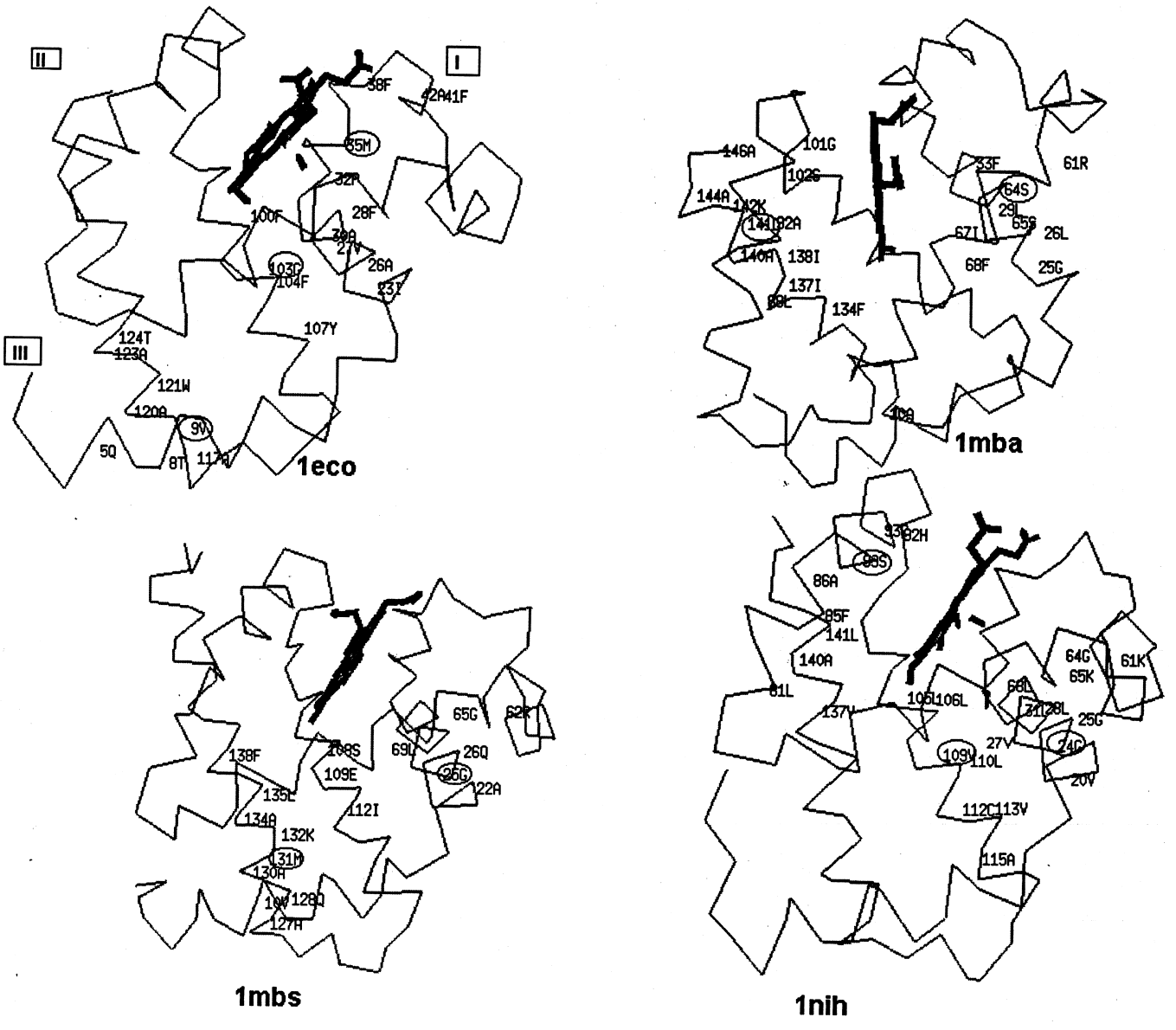

Fig. 3. A graphical representation (generated by VMD [30]) of core clusters (with center marked by O) for the globin family.

amino acid residues important in stabilizing the folding intermediates. Recently we have carried out [19] similar clustering analysis on the side chains of amino acids in proteins. The present analysis of backbone clusters can complement the results of side chain cluster analysis.

4. Summary

The graph theoretical approach has been used to identify non-bonded backbone clusters in proteins. The non-covalent interactions among spatially neighboring residues in proteins are considered. The adjacency matrix is constructed by considering non-sequential residues appearing within a distance of 6.5 Å of the C_α atom of the chosen residue. The diagonalization of the matrix yields eigenvalues and eigenvectors, which can be related to backbone clusters in protein structures and the center of clusters, respectively. The method has been applied to the proteins of the RNase-A and globin family. Three clusters in similar regions in many of the RNase-A and family proteins are identified. Several important conserved residues are part of the clusters with high vector components. The globin family proteins have been identified to be made up of approximately three clusters located around the porphyrin. The helix–helix crossing regions containing generally the small residues such as glycine, alanine and serine form the center of clusters with high branching parameters as found by eigenvector components.

The cluster identification method developed here is simple, straightforward and gives consistent results. The method can prove to be useful in detecting structural similarities in proteins. The method can also be useful in identifying domains in proteins unambiguously. The current investigation has the potential to assist experiments such as mutational studies designed to investigate the process of protein folding and in homology modeling of protein structures.

Acknowledgements

We are thankful to the Super Computer Education and Research Center, Indian Institute of Science, Bangalore, India for the computational facility. Swarna Mayee Patra thanks the Department of Biotechnology, India for a Post Doctoral fellowship. We thank Latha for her help during preparation of the manuscript and N. Kannan for useful discussions

References

[1] R.L. Baldwin, Nature (London) 346 (1990) 409–410.
[2] P.S. Kim, R.L. Baldwin, Annu. Rev. Biochem. 59 (1990) 631–660.
[3] J. Heringa, P. Argos, J. Mol. Biol. 220 (1991) 151–171.
[4] K.A. Dill, S. Bromberg, K. Yue et al., Protein Sci. 4 (1995) 561–602.
[5] M.M. Gromiha, S. Selvaraj, Biophys. Chem. 77 (1) (1999) 49–68.
[6] F.M. Richard, Ann. Rev. Biophys. Bioeng. 6 (1977) 151–176.
[7] Z. Dosztanyi, A. Fiser, I. Simon, J. Mol. Biol. 272 (1997) 597–611.
[8] R. Sowdhamini, T.L. Blundell, Protein Sci. 4 (1995) 506–520.
[9] M.H. Zefus, Protein Sci. 4 (1995) 1188–1202.
[10] M.B. Swindells, Protein Sci. 4 (1995) 93–102.
[11] D.R. Flower, Protein Eng. 7 (11) (1994) 1305–1310.
[12] I. Koch, F. Kaden, J. Selbig, Proteins 12 (4) (1992) 314–323.
[13] P.J. Artymiuk, D.W. Rice, E.M. Mitchell, P. Willett, Protein Eng. 4 (1) (1990) 39–43.
[14] T. Haliloglu, I. Bahar, B. Erman, Phys. Rev. Lett. 79 (16) (1997) 3090–3093.
[15] I. Bahar, A.R. Atilgan, R.L. Jernigan, B. Erman, Proteins Struct. Funct. Genet. 29 (1997) 172–185.
[16] M.C. Demirel, A. Atilgan, R.L. Jernigan, B. Erman, I. Bahar, Protein Sci. 7 (1998) 2522–2532.
[17] I. Bahar, A.R. Atilgan, B. Erman, Fold Des. 2 (1997) 173–181.
[18] S.M. Patra, S. Vishveshwara, Int. J. Quantum. Chem. 71 (1999) 349–356.
[19] N. Kannan, S. Vishveshwara, J. Mol. Biol. 292 (2) (1999) 441–464.
[20] Encyclopedia of Computational Chemistry, in: P.V.R. Schleyer (Eds.), vol. 2, John Wiley and Sons, NY, Singapore 1998 p. 1174.
[21] S.K. Panjikar, M. Biswas, S. Vishveshwara, Acta Crystallogr. D53 (1997) 627–637.
[22] S. Miyazawa, R.L. Jernigan, Macromolecules 18 (1985) 534–552.
[23] S. Miyazawa, R.L. Jernigan, J. Mol. Biol. 256 (1996) 623–644.
[24] L. Hagen, A.B. Kahng, IEE Trans. Comput. Aided Des. III (9) (1992) 1074–1084.

[25] I. Gutman, D. Cvetkovic, Croat. Chem. Acta 49 (1977) 115–121.
[26] M. Randic, J. Chem. Inf. Comput. Sci. 15 (1975) 105.
[27] C.M., Cuchillo, M., Vilanova, M.V., Nogues, Ribonucleases Structures and Functions, in: G.D. Alessio, J.F. Riordan (Eds.), copyright by Academic Press 1997, p. 271.
[28] D. Juminaga, W.J. Wedemeyer, G. Juarez, R. McDonald, M.A. H.A. Scheraga, Biochemistry 36 (1997) 10131–10145.
[29] A.M. Lesk, C.J. Chothia, Mol. Biol. 136 (1980) 225–270.
[30] W. Humphrey, A. Dalke, K. Schulten, Visual molecular dynamics, J. Mol. Graph. 14 (1) (1996) 33–38.