PROTEINS: Structure, Function, and Bioinformatics 59:616–626 (2005)

Identification of Domains and Domain Interface Residues in Multidomain Proteins From Graph Spectral Method

Ramesh K. Sistla, Brinda K. V., and Saraswathi Vishveshwara
Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India

ABSTRACT **We present a novel method for the identification of structural domains and domain interface residues in proteins by graph spectral method. This method converts the three-dimensional structure of the protein into a graph by using atomic coordinates from the PDB file. Domain definitions are obtained by constructing either a protein backbone graph or a protein side-chain graph. The graph is constructed based on the interactions between amino acid residues in the three-dimensional structure of the proteins. The spectral parameters of such a graph contain information regarding the domains and subdomains in the protein structure. This is based on the fact that the interactions among amino acids are higher within a domain than across domains. This is evident in the spectra of the protein backbone and the side-chain graphs, thus differentiating the structural domains from one another. Further, residues that occur at the interface of two domains can also be easily identified from the spectra. This method is simple, elegant, and robust. Moreover, a single numeric computation yields both the domain definitions and the interface residues. Proteins 2005;59:616–626.** © 2005 Wiley-Liss, Inc.

Key words: protein domain; domain interface; protein structure graph; eigen spectra; vector component

INTRODUCTION

Multidomain proteins have been defined and identified variously after Richardson[1] defined them as subunits of the polypeptide chain, which form an independent stable folded structure. Almost all approaches agree that the number of connections formed within a domain is more than those formed across domains and the identification algorithms try to exploit this feature in assignment of domains.

One approach for identifying domains has been the detection of hydrophobic core in a protein. Swindells' DETECTIVE algorithm considered three requirements to identify hydrophobic cores. They are based on the secondary structure, the side chain accessibility, and the side-chain contacts.[2] Another approach exploits a more obvious feature of the domains, which is, that there are more numbers of connections within a domain compared to those across domains. This feature has been exploited in the FSSP database of Holm and Sander,[3] DOMAK database of Siddiqui and Barton,[4] and by Islam et al.[5] A database presented by Sowdhamini et al.[6] identifies the domains based upon the clustering of secondary structural elements.

Jones et al.[7] use a consensus approach for domain identification in their CATH database. Four domain assignment methods, namely, PUU (parser for protein unfolding units),[3] DETECTIVE,[2] and DOMAK,[4] and a method by Islam et al.[5] are used in this approach. If all four methods are unanimous on the identification of the domains in a particular protein, then the domains are automatically detected. If there is a difference among these methods, manual judgment is made about the best definition among the four. SCOP[8] is another exhaustive database in which the proteins are checked manually for evolutionary relationship, function, etc. before the entry is made in the database. Veretnik et al.[9] have recently published an exhaustive comparison of various algorithmic and expert methods of domain assignment and commented on the inconsistencies present in various methods. They have also listed out the central issues that have to be addressed for arriving at a consistent definition of structural domains. Ying Xu et al.[10] have reported the automatic decomposition of multidomain proteins in to individual domains by a graph theoretic approach. They have implemented the algorithm as a computer program called Domain Parser.

With the rapid rise in the number of entries in the Protein Data Bank (PDB),[11] it is highly desirable to have a simple, straightforward, numerically robust, and computationally sound method to automatically identify the different domains in proteins. In this paper, we have proposed a simple, single numeric and completely automatic computation for domain identification. This method is based on graph theoretic technique, which considers the overall connectivity and topology of the protein structure. It exploits the feature that the interactions between the amino acid residues are higher within a structural domain than across domains. Our methodology takes the atomic coordinates of the protein as the input and identifies major subclusters as domains of a connected protein graph.

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

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

Received 19 June 2004; Accepted 1 December 2004

Published online 23 March 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20444

Furthermore, the method also gives the domain–domain interface residues as a solution to the connectivity matrix obtained from the covalent and noncovalent interactions in the protein. We believe that this is a unique method of domain and the interface residue identification, which is quantitative as well as amenable to automation. In addition, the domain information is automatically incorporated into the protein coordinate file, which helps in the visualization of different domains in different colors. The methodology is described below, which is followed by the Results and Discussion section.

METHOD

The protein molecule is represented as a graph of connected amino acid residues. The protein graph is constructed by considering each amino acid in the protein as a node. Two types of definitions are used to obtain the connections (edges) among the nodes. One method considers the C^{α} atom distances between residues to define the edges, a criterion that we had earlier used in the context of the identification of the backbone clusters in proteins.[12] In this formalism, the protein backbone graph (PBG) was constructed by considering the C^{α} atom of each residue in the protein as a node and any two C^{α} atoms that are at a distance less than 6.5 Å are connected by an edge. The protein backbone graph can be represented as an $N \times N$ (N = number of residues in the protein = number of nodes in the graph) Laplacian matrix as follows:

$$L_{ij} = -1, \text{ if } R_{ij} \leq R_c \text{ and } i \neq j \quad (1)$$

$$L_{ij} = 0, \text{ if } R_{ij} > R_c \text{ and } i \neq j \quad (2)$$

$$L_{ij} = -\sum_i L_{ij}, \text{ for } i = j \quad (3)$$

where R_c is the cutoff distance and is usually taken as 6.5 Å and i goes from 1 to N, the number of residues in the protein.

In the second method, the protein side-chain graph (PScG) was constructed on the basis of the details of the side-chain interactions, which is quantified in terms of the extent of interaction as given by Kannan and Vishveshwara.[13] A brief description of this method is given here.

The interaction criterion between two residues is evaluated based on the number of pairwise atom–atom contacts that occur between the side-chain atoms of the two residues in the three-dimensional structure of the protein. Any two side-chain atoms belonging to two residues, which come within a distance of 4.5 Å is considered as a side-chain atom–atom contact. The number of such contacts between the two residues is counted and the percentage contact is evaluated by comparing with the normalized values obtained from a nonredundant dataset.

Thus, the interaction criterion is evaluated as percentage given below:

$$I_{ij} = (n_{ij}/N_i) \times 100 \quad (4)$$

where, I_{ij} is the interaction between residues i and j. n_{ij} is the number of side-chain atom–atom contacts (< 4.5 Å between residues i and j). N_i is the normalization value for residue-type i. N_i has been evaluated from a nonredundant set of protein structures and is based upon the maximum number of interactions (in terms of atom–atom contacts of < 4.5 Å) that a residue-type normally makes in protein structures. The normalization values for the 20 amino acids are given by Kannan and Vishveshwara[13] and reproduced in Table I.

TABLE I. Normalization Values Used for Clustering Residues into Domains

Amino acid	Normalization values
Ala	55.76
Arg	93.79
Asn	73.41
Asp	75.15
Cys	54.95
Gln	78.13
Glu	78.83
Gly	47.31
His	83.74
Ile	67.95
Leu	72.25
Lys	69.61
Met	69.26
Phe	93.31
Pro	51.33
Ser	61.39
Thr	63.71
Trp	106.70
Tyr	100.72
Val	62.37

The interaction criterion is evaluated for all pairs of nxn residues (ij residue pairs) in the protein. Any two residues, which have interaction greater than a specified value (interaction cutoff, I_{cutoff}), are connected by an edge in the graph. Thus, we get a connected protein graph. This graph is then represented as a Laplacian matrix as follows:

$$L_{ij} = -1, \text{ if } i \neq j, \text{ i and j are connected} \quad (5)$$

$$L_{ij} = 0, \text{ if } i \neq j, \text{ i and j are not connected} \quad (6)$$

$$L_{ij} = -\sum_i L_{ij}, \text{ for } i = j, (i = 1 \text{ to } N, \text{ where N is the number of residues in the protein}) \quad (7)$$

The Laplacian matrices thus obtained from PBG and PScG are then diagonalized to obtain the eigen spectra. It is known[14] that the eigenvector components corresponding to the second lowest eigenvalue contain information about clusters present in a graph. For the sake of brevity, these eigenvector components are denoted as 2evc henceforth. All the nodes that belong to a particular cluster have the same magnitude of 2evc. The values of 2evc are sorted along with the corresponding nodes. Thus the nodes (residues in case of proteins) forming a cluster can be identified from 2evc.

The above methodology was used earlier[13] to identify distinct disjoint clusters, which arise due to high interac-

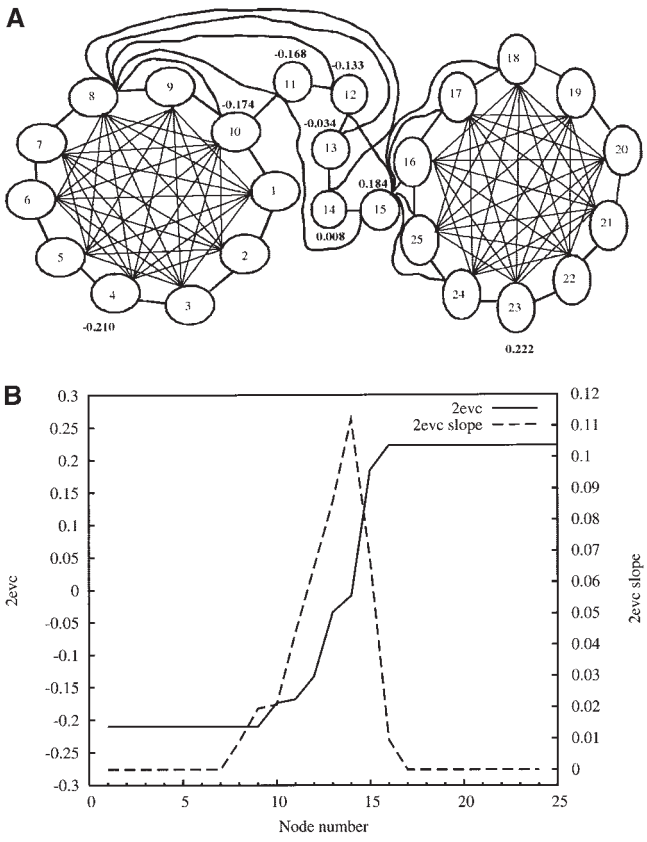

Fig. 1. **A**: A 25-node graph. Its connectivities and the eigenvector components (2evc) of individual nodes corresponding to the second lowest eigenvalue. **B**: 2evc and smoothened 2evc slope plot for the 25-node example graph shown in **A**.

tion cutoff (about 12%). However, a reduction in the cutoff to about 4% yields the clusters of the size of domains observed normally in proteins. Furthermore, we have considered the whole protein as a single cluster by connecting all the sequentially neighboring residues. In the present study, we show that the protein domains form subclusters of the protein graph in both PBG and PScG by applying the same principle, which was used earlier for cluster detection.[12,13] This concept is made clear with a simple example as shown below.

The example graph shown in Figure 1(a) consists of 25 nodes. The Laplacian matrix was constructed for this graph as defined earlier and was diagonalized. Figure 1(b) shows the plot of the sorted 2evc versus the node number. Each plateau in this figure represents a subcluster or a domain. A small number of nodes connecting both the subclusters have vector component values in between those of the two plateaus. The real protein cases may not be as simple as this example. The plateau region may not be easily demarcated from the 2evc plot. However this problem can be addressed by considering the slope at each point as in the 2evc-slope plot shown in Figure 1(b). There is no significant change of slope in a plateau region and there is a sharp change in slope in the interface region between the domains. The domain and the interface regions can be clearly identified from this slope plot. The

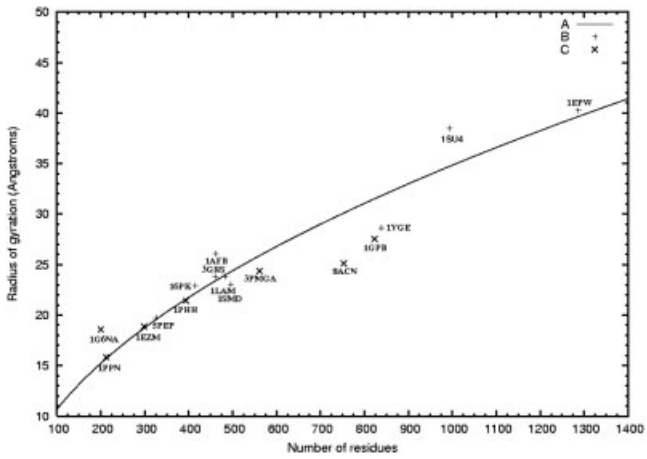

Fig. 2. Plot of number of residues versus radius of gyration for (**A**) nonredundant dataset of 285 proteins (best fit as per equation $R_G = N^{0.511}$), (**B**) Proteins we have chosen for our study, (**C**) Proteins chosen from Xu et al.[10]

interface residues fall in the breadth of the peak and the plateaus on either side of the peak correspond to two domains. Since the graph spectra of real proteins may be complicated, it may become difficult to identify the domains and the domain interfaces from the 2evc plot. However, they can be better resolved from the 2evc-slope plot.

Compactness of protein structures can vary widely. A single default value for the selection of the edge in the PSGs may not be able to take into account the variations of the compactness in different proteins. However, this can be addressed by employing a variable cutoff criterion for the formation of the Laplacian matrix. Since the radius of gyration (R_G) of a protein is a measure of its compactness, we select out cutoff value for edge formation by normalizing it with respect to R_G as given below. R_G was calculated from the C^α coordinates of the proteins by the formula:

$$R_G = \frac{\sqrt{\sum_{i=1}^{N} R_i^2}}{N} \qquad (8)$$

where R_i is the distance of the C^α atom of the i^{th} residue from the centroid of the protein and N is the number of residues in the protein.[15]

We calculated the R_G for a nonredundant set of 285 proteins. The best fit of the plot of N versus R_G for this dataset was governed by the equation $R_G = N^{0.511}$. This plot is shown in Figure 2, which also has the R_G of the proteins we have chosen (depicted by + and ×) for domain analysis. We adopt a default value of 4% and 6.5 Å cutoffs respectively in PScG and PBG if the R_G value of the selected protein is close to the curve given in Figure 2. However, if the protein is significantly above the curve (less compact structure), a cutoff of $< 4\%$ for PScG and > 6.5 Å for PBG is recommended. Similarly, if the protein is significantly below the curve (very compact structure), a cutoff of $> 4\%$ for PScG and < 6.5 Å for PBG is recom-

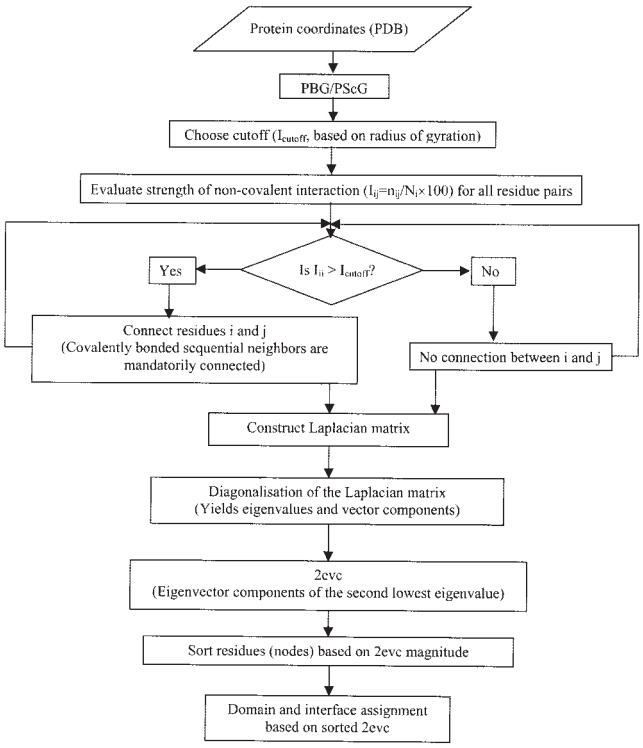

Fig. 3. Flow-chart depicting the graph spectral algorithm used for domain decomposition.

mended. While the choice of a cutoff value by this method might need fine tuning, there is no doubt that it can serve as a very effective guide for domain identification in diverse proteins.

Furthermore, we have automated the visualization of different domains by incorporating the relevant information into the PDB file of protein coordinates. The columns 61–66 in a PDB file denote the experimental B-factors. Our program replaces this column by the scaled 2evc for all the atoms in the protein molecule. On visualization, the B-factor coloring scheme is used, wherein all the residues in a particular domain having similar vector components take up the same color. Thus, different domains assume different colors due to differences in the magnitude of their vector components. The interface region between the two domains takes up a color gradation ranging from the color of one domain to that of the other. We have used VMD[16] package in our present study for this purpose.

The program to carry out the domain identification and to modify the PDB for domain display has been written in FORTRAN and makes use of various shell commands of Linux and the visualization software such as VMD. The sequence of steps followed for the domain and interface identification is presented as a flow chart in Figure 3. This algorithm has been tested on a large number of proteins. Here, we present the results pertaining to a few proteins (nine), which have been selected on the basis of diverse domain organization and elucidate various aspects of our methodology. Additionally, we have also discussed seven proteins taken from Xu et al.[10], which were extensively analyzed for the evaluation of domain assignment methods. The performance of the current spectral method of identifying the domains using the PBG and the PScG in comparison with the known methods have been presented for the sixteen selected proteins, which are described in the next section.

RESULTS AND DISCUSSION

As explained in the methods section, we have used two different types of representations in this analysis, namely the protein backbone graph (PBG) and the protein side-chain graph (PScG) so as to obtain insights into the organization of structural modules called domains within the protein structure. The significant results of this analysis pertaining to domain and interface residue identifications are summarized below.

Identification of Domains from PBG and PScG

The methods section gives in detail the role of graph spectral parameters in domain identification. Here, we present the domain definition results obtained using these parameters on nine different proteins. Table II gives the number of residues in nine of the selected proteins, the definition of different domains as per SCOP, CATH, DOMAIN PARSER, and our study using PBG and PScG. Table III shows the number of domains assigned by these five methods in the seven proteins selected from Xu et al.[10] Tables II and III show that there is significant correlation in the domain definitions among the five methods although there are discrepancies in a few cases. These are discussed in detail in the next section. The method of identification of domain definitions using the present method in the nine protein cases can be understood from Figure 4, which shows the plot between 2evc and the node numbers for all the nine proteins. The solid line in these plots indicates the results obtained from the PBG and the dotted line indicates the results obtained from the PScG. The domain definitions in these proteins are identified based on the closeness in the magnitudes of 2evcs of the residues belonging to a domain. As explained earlier, the present method uses the feature that the connections among residues are higher within a domain than across domains. This factor is reflected in the magnitude of 2evcs of the residues, which is obtained from the diagonalization of the PBG and PScG. The residues belonging to a domain have 2evcs that are closer in magnitude than to those belonging to different domains. Hence, the 2evc plot gives the domain demarcations and the residues belonging to each domain in the protein structure. We also find that the PBG and PScG correlate well with regard to domain definitions in most of the cases. The differences seen in the definitions of domains by PBG and PScG as observed in a few cases arise due to the fact that the PBG takes into account only the C^{α} interactions among residues within a protein structure, whereas the PScG takes into account all the side-chain atom interactions. Hence, PScG is more rigorous than PBG, however, PBG is much simpler to compute. Figure 5 shows the three-dimensional representations of the nine proteins and their domain definitions according to PScG.

TABLE II. Domain Definitions Obtained From the C^{α} and Side-Chain Representations and Comparison with SCOP, CATH, and DOMAIN PARSER Definitions

Protein	PDB Code	nres	Number of domains and residues in different domains									
			SCOP		CATH		Domain Parser		PBG		PScG	
Lipoxygenase	1YGE	839	2	1–149, 150–839	5	9–167, 168–267, 268–356, 357–490, 491–839	2	1–145, 146–839	2	1–158, 159–839	2	1–123, 124–839
Lipoxygenase (@12% Cutoff PScG)	1YGE	839	—	—	—	—	—	—	—	—	5	1–114, 115–265, 266–351, 352–478, 479–839
Leucine aminopeptidase	1LAM	484	2	1–159, 160–484	2	1–165, 166–483	2	1–162, 163–484	2	1–158, 159–484	2	1–72, 73–484
Amylase	1SMD	495	2	1–403, 404–495	2	2–403, 404–496	4	2–254, 255–341 & 383–399, 342–382, 400–496	2	1–403, 404–484	2	2–399, 400–496
Glutathione reductase	3GRS	461	2	18–165 & 291–363 & 166–290, 364–478	3	18–160 & 290–365, 161–289, 366–478	4	18–64 & 106–159 & 291–364, 65–105, 160–290, 365–478	3	18–162 & 290–363, 163–289, 365–478	2	55–112 & 160–291 & 369–478, 18–54 & 113–159 & 292–368
Pepsin	5PEP	326	1	1–326	2	1–170, 171–327	2	1–171, 172–327	2	1–193, 194–327	2	1–117, 118–327
Calcium ATPase (2% cutoff PScG)	1SU4	994	4	125–239, 344–360 & 600–750, 361–599, 1–124 & 240–343 & 751–994	4	1–43 & 124–242, 44–123 & 243–345 & 746–994, 359–602, 346–358 & 603–745	5	1–16 & 147–241, 358–602, 323–357 & 603–751, 17–62 & 101–146 & 242–262, 63–100 & 263–322 & 752–994	4	1–41 & 106–235, 42–105 & 236–328 & 751–994, 329–361 & 600–750, 362–599	4	1–44 & 115–252, 45–114 & 245–318 & 755–994, 366–597, 316–365 & 598–754
Phosphoglycerate kinase	16PK	415	2	—[b]	2	5–192, 199–406	2	5–204 & 407–419, 205–406	2	1–202, 203–415	2	5–202, 203–407
Clostridium neurotoxin	1EPW	1287	4	1–533, 534–861, 862–1079, 1080–1287	4	1–533, 534–861, 862–1079, 1080–1287	4	1–533, 534–861, 862–1077, 1078–1290	4	1–532, 533–858, 859–1080, 1081–1287	4	1–523, 524–850, 851–1090, 1091–1287
Mannose binding lectin-A (@ 6.5 Å Cutoff PBG)	1AFB[a]	462	2	74–103 of A,B,C, 104–230 of A,B,C	3	73–226A, 73–226B, 73–226C	3	97–226 of A, B, C	3	73–226A, 73–226B, 73–226C	2	99–226 A,B and C, 73–98 A, B and C
Mannose binding lectin-A (@ 7.5 Å Cutoff PBG)	1AFB*	462	—	—	—	—	—	—	3	73–96 A, B & C, 97–226 B & C, 97–226A	—	—

[a]Trimer, comprising of A, B, and C chains.
[b]SCOP classifies 16pk as a single-domained protein consisting of two similar domains.
Domains are separated by a comma (,) and discontinuous regions of a domain are separated by the symbol "&" in all the entries in this table.

TABLE III. Domain Definitions of Selected Proteins From Xu et al.[10] by Various Methods

PDB	Protein	SCOP	CATH	Domain parser	PScG	PBG
1GPB	Glycogen phosphorylase	1	2	5	1(4%), 2(8%)	2(6.5 Å & 7.5 Å)
1PPN	Papain	1	1	1	2	2(6.5 Å), 1(8.0 Å)
8ACN	Aconitase	2	4	2	2(4%), 4(16%)	2(6.5 Å), 4(4.5 Å)
1PHH	p-hydroxybenzoate hydroxylase	2	2	3	2	2
1G6N	Catabolic gene activator protein	2	2	1	3(4%), 2(2%)	3(6.5 Å), 2(8.0 Å)
3PMGA	Phosphogluco-mutase	2	4	4	2(4%), 3(2 & 0%)	2(6.5 Å), 4(8.0 Å)
1EZM	Elastase	1	2	2	2	2

The figure also shows the SCOP/CATH definitions of the nine proteins for the sake of comparison. The domains identified by SCOP/CATH are shown in cartoon representation with a different color for each domain. The domains identified by our definitions are shown in Van der Waal's (VDW) representation with an automated graded coloring scheme based on the magnitude of the 2evc. It is evident from Figure 5 that all the atoms in a domain have comparable eigenvector components and hence they are automatically displayed in the same color. A figure for the domain definitions of the nine chosen proteins obtained by PBG, which is similar to that of PScG (Fig. 5), is given as a supplementary material (Fig. S1).

In the present study, the PScG representation makes use of the normalization values for each amino acid type (Table I). This takes into account the size and the property of the residue. This information is built into the Laplacian matrix of PScG and hence the results obtained are based on rigorous packing considerations. Although the packing density in most proteins is very similar, there are a significant number of cases with high and low densities than the average value. As explained in the methods sections, the packing is approximately evaluated using the radius of gyration. Figure 2 shows that the packing density of the proteins 1G6NA, 1AFB, 3PGK, and 1SU4 are significantly higher whereas that of 3PMGA, 8ACN, 1GPB,

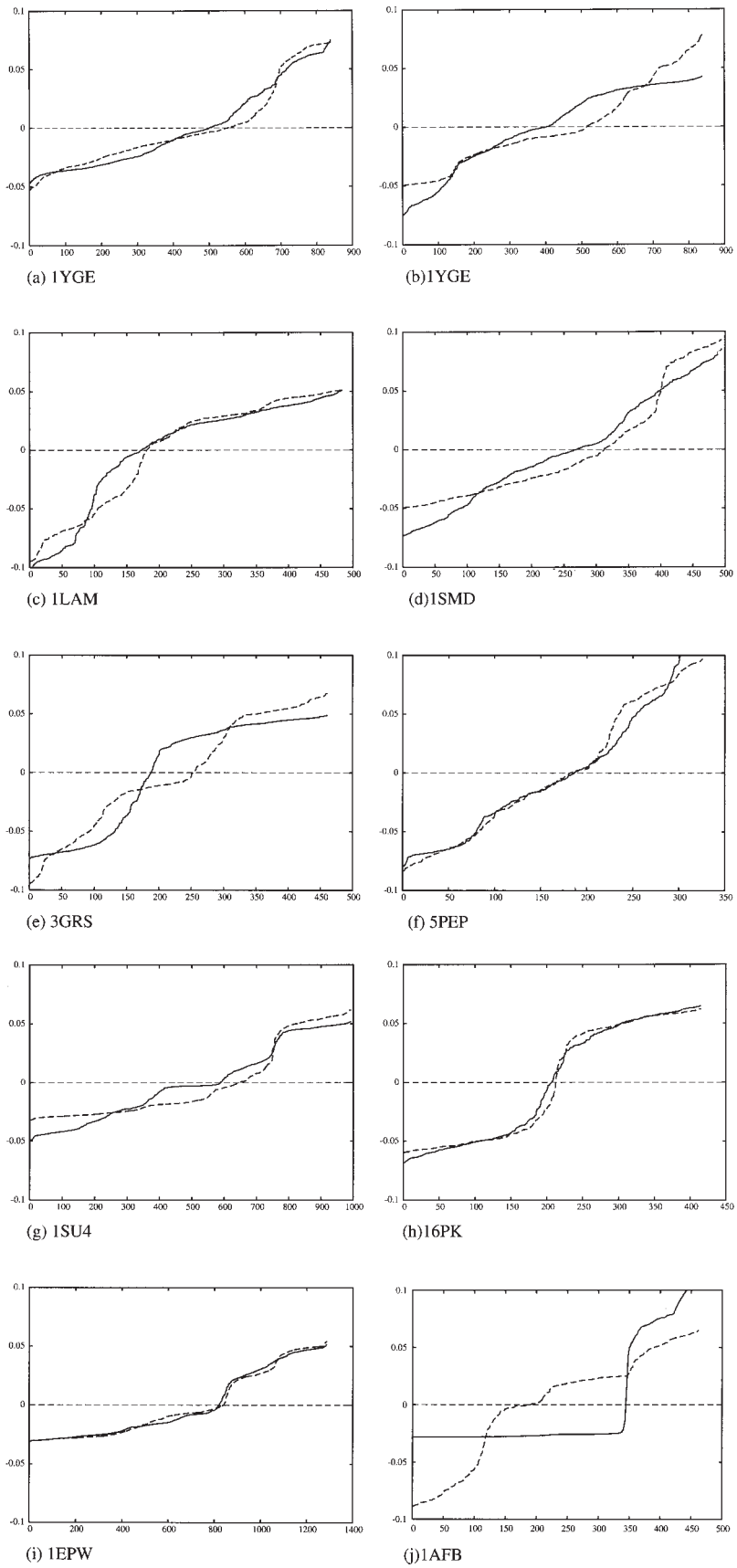

Fig. 4. 2evc plots for different proteins with sorted 2evc on the Y axis and the node number on the X axis. The solid line indicates PBG representation and dashed line indicates the PScG representation. (**a**) 1YGE (4% PScG cutoff and 6.5 Å PBG cutoff), (**b**) 1YGE (8% PScG cutoff and 5.5 Å PBG cutoff), (**c**) 1LAM, (**d**) 1SMD, (**e**) 3GRS, (**f**) 5PEP, (**g**) 1SU4, (**h**) 16PK, (**i**) 1EPW, (**j**) 1AFB.

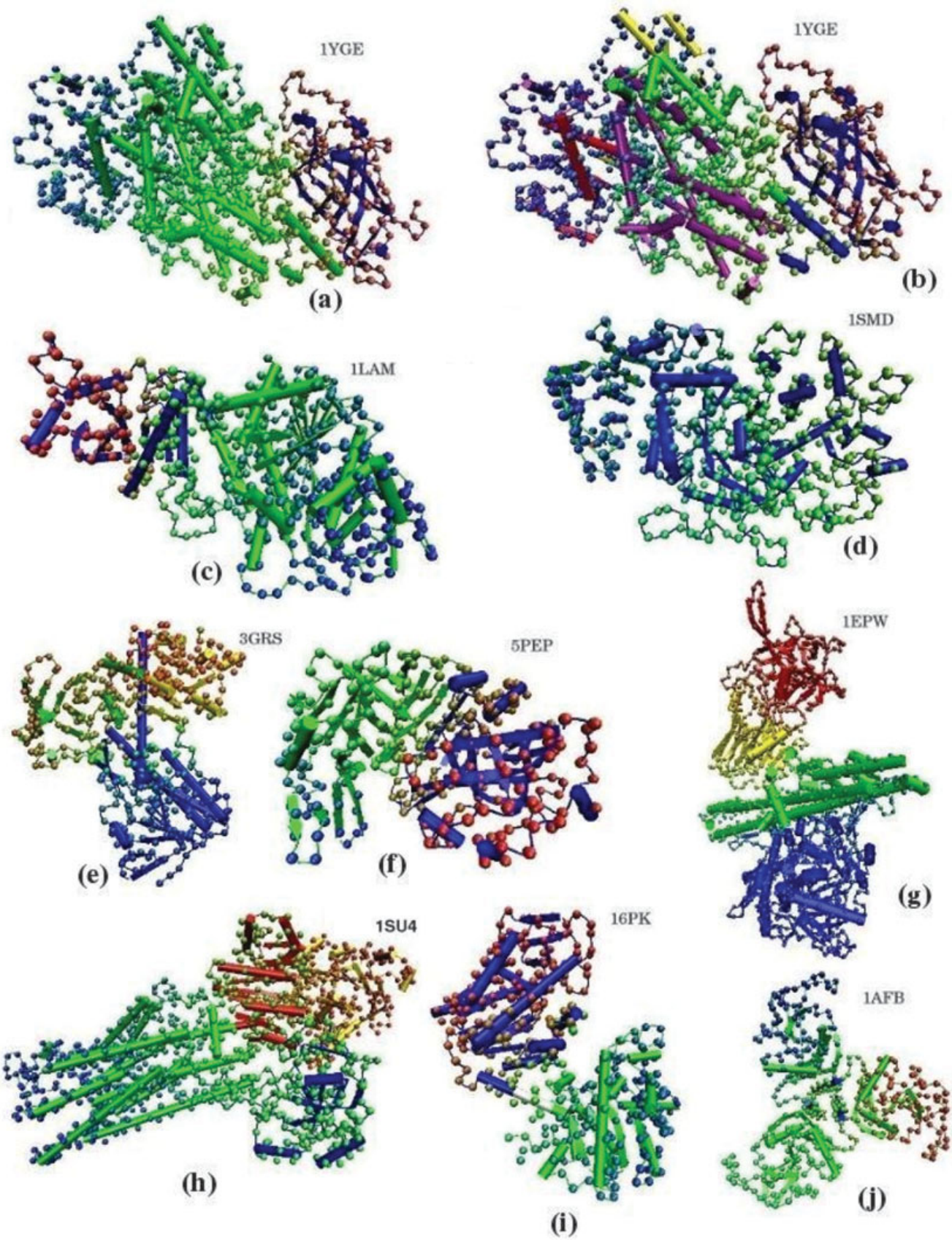

Fig. 5. Visualization of different domains for different proteins, obtained from the PScG representation. The cartoon representations show the SCOP/CATH definitions while the van der Waal's representation shows the definitions obtained from the present method using PScG. (**a**) 1YGE (4% cutoff and comparison with SCOP definition), (**b**) 1YGE (8% cutoff and comparison with CATH definition), (**c**) 1LAM, (**d**) 1SMD, (**e**) 3GRS, (**f**) 5PEP, (**g**) 1SU4, (**h**) 16PK, (**i**) 1EPW, (**j**) 1AFB.

and 1YGE are lower than the best-fit curve and the others fall very close to the curve. We observe that a 4% interaction cutoff for PScG and 6.5 Å cutoff for PBG are optimal for domain identification of proteins which fall close to the best-fit curve in Figure 2. However, those cases where the points are above and below the curve, need to be handled carefully by using varied cutoffs as explained earlier. A lower PScG cutoff (0–3%) and a higher PBG cutoff (7.0–8.0 Å) needs to be used in case of proteins that are less compact than expected (above the curve) so that the connections within the domains can be maximized as in the case of 1G6NA (Table III). On the contrary, a higher PScG cutoff (6–16%) and a lower PBG cutoff (4.5–5.5 Å) needs to be employed in case of proteins that are more compact than expected (below the curve) so that the connections across the domains can be minimized as in the case of 1YGE (Table II). However, this is only a general guideline and there are some exceptions to this cutoff rule as seen in 1PPN and 3PMGA (Fig. 2, Table III). For instance, though 1PPN falls exactly on the curve in Figure 2, the cutoffs had to be varied from the optimal ones to assign domains correctly. 3PMGA is a different case where, instead of using a cutoff (higher for PScG and lower for PBG) as recommended for proteins falling below the curve in Figure 2, the opposite had to be done to obtain the proper domain definitions. Thus, the radius of gyration plot gives only a broad idea regarding the cutoff that needs to be employed for domain identification, however it works for most cases. Hence, we find that analyzing PScG and PBG at different cutoffs can give useful insights about the protein structure and domain separation.

Comparison of the Present Method with SCOP, CATH, and DOMAIN PARSER

By and large different methods agree reasonably well in assigning both the number of domains as well as the number of residues in a domain. However, there are several cases of disagreement too. The proteins in Tables II and III have been particularly selected to represent samples of good agreement and of disagreeing cases. As can be seen from Tables II and III, SCOP, CATH, and DOMAIN PARSER differ among themselves in the domain definitions of all cases except 1EPW, 1LAM, 16PK, 3PGK, and 1PPN. Among these proteins, our method disagrees with the other three significantly only in the case of 1PPN, where the agreement with SCOP, CATH, and DOMAIN PARSER (Table III) is seen only by PBG at 8.0 Å cutoff. In all the other examples considered in this dataset, PScG and PBG tally either with SCOP or CATH or both. In the examples where SCOP, CATH, and DOMAIN PARSER disagree, SCOP differs in its domain assignment since the function and fold of the protein domains are emphasized more rather than structural modules, whereas the other methods classify on a purely structural basis. Let us consider a few specific examples, which elucidate the correlations and discrepancies among all the five methods considered in this analysis.

Clostridium neurotoxin (1EPW) is one of the largest proteins that we have considered and is a good example of a case where all the five domain assignments correlate very well (Table II). It also falls very close to best-fit curve in Figure 2. It has 1287 residues and four domains as classified by both SCOP, CATH, and DOMAIN PARSER. We get the same classification as the other three from both PScG [Fig. 5(g)] and PBG, indicating that the present method works for large proteins as well. p-Hydroxybenzoate hydroxylase (1PHH) has two domains as classified by SCOP, CATH, PScG, and PBG. However, DOMAIN PARSER assigns three domains to this protein (Table III). This protein falls exactly on the best-fit curve in Figure 2, which indicates that its compactness is as expected and hence the normal cutoff of 4% and 6.5 Å works very well in this case.

Lipoxygenase (1YGE) is classified as having two domains in SCOP, five domains in CATH, and two domains in DOMAIN PARSER (Table II). Since this protein falls below the curve in Figure 2, a different cutoff is needed in this case. PScG at 4% cutoff gives two domains in our study [Fig. 5(a)]. By increasing the strength of the cutoff to 8%, we get five domains, which is the same as CATH classification [Fig. 5(b)]. This is because, by increasing the cutoff, we are eliminating the weak interactions throughout the protein, including those at the domain–domain interface regions. As a result, the not-so-well-resolved domains at 4% cutoff gets resolved into five domains when cutoff is raised to 8%. The same result is obtained in PBG representation by decreasing the R_c cutoff distance from 6.5 Å to 5.5 Å.

Another example that illustrates the correlation of compactness with the cutoff value for edge formation is the protein Mannose binding Lectin-A [1AFB, Fig. 5(j)], which falls well above the curve in Figure 2. 1AFB comprises of three chains. As per SCOP, the sequences 74–103 of each of the three chains form a coiled coil, while the rest of the sequence in each chain is a different structural domain, but has the same C-type lectin fold. Hence, this is classified by SCOP as a two-domain protein with the triple coiled coil as one domain and the rest as three copies of the C-type lectin domain. CATH and DOMAIN PARSER classify each chain as a separate domain and hence this is a three-domain protein as per both (Table II). We observe the following. Figure 5(j) shows the PScG representation of 1AFB, which gives the same result as SCOP. The PBG of 1AFB at $R_c = 7.5$ Å, gives the triple coiled coil as a separate domain. When the PBG was constructed using the $R_c = 6.5$ Å, we get the same classification as CATH. However, when this cutoff distance was increased to 7.5 Å, we obtain the same classification as SCOP.

Figure 5(h) shows the protein Calcium ATPase (1SU4) in PScG representation. This protein shows discontinuous segments in the sequence as part of the same domain (Table II). We are able to identify the discontinuous segments in a domain clearly, thus underscoring the importance assumed by topological connections in our method. Moreover, this protein has four domains as assigned by SCOP, CATH, PScG, and PBG. However, DOMAIN PARSER disagrees with the others and assigns five domains to it as can be seen from Table II. Since 1SU4 falls

significantly above the curve in Figure 2, the cutoffs were varied accordingly while employing PScG and PBG.

To summarize, Tables II and III give an overview of how the five domain classification methods namely SCOP, CATH, DOMAIN PARSER, PScG, and PBG compare with each other. We observe that PScG and PBG correlate very well with both SCOP and CATH and in cases where there are discrepancies between SCOP and CATH, we are able to reproduce both using variable cutoff. We have also presented a method for choosing the right cutoff based on the compactness of the protein using radius of gyration as a general measure of the same.

Subsequent to the completion of the current work, Veretnik et al.[9] reported their findings on domain definitions by various methods. This paper cites some examples of proteins (1AVHA, 1TAHB, 5FBP, 1GPB, 1SMR, and 1CHRA) for which there is no consensus on domain definitions between the established methods. On examining these proteins by our method, we find that the results obtained from both PBG as well as PScG representation agree very well with CATH and to a good extent with AUTHORS (This is an annotation given by Veretnik et al.[9] for domain definitions given by the authors of the respective crystal structure papers as compiled by Islam et al.[5]).

Identification of Domain Interface Residues

Understanding the nature of the domain interfaces is important from the perspective of protein structure and function because the interactions between domains are modes of communication within the protein and are necessary for protein function and regulation. Moreover, the flexibility and mobility between the protein domains also have functional implications. Therefore, various research groups have extensively carried out the identification of domain interfaces and their analyses. Jones et al.[17] have analyzed the properties of the domain interfaces such as polarity, planarity, amino acid composition, etc. They made a comparison of these parameters between the domain interfaces and protein subunit interfaces and found that the domain–domain interfaces are similar to subunit interfaces in many ways. Such an analysis has aided in the correlations of protein structure and function.

We present a different method of identifying domain interfaces in this paper as can be seen from Figure 6. Figure 6(a) shows the 2evc plots of the protein phosphoglycerate kinase (16PK) obtained using PScG. The two plateau regions seen in Figure 6(a) correspond to the two structural domains in the proteins and the residues falling under the peak between the plateaus in the 2evc smoothened slope plot [also shown in Fig. 6(a)] correspond to the interface residues between the two domains. The domains and the interface residues thus identified from Figure 6(a) are mapped onto to the protein structure as shown in Figure 6(b). The two domains are shown in cartoon representation and the interface residues are in VDW representation. It is clear from Figure 6(a,b) that the residues identified from Figure 6(a) actually form the interface between the two domains of the protein. Thus, the domain interface residues can be easily identified using the present method. The domain interfaces identified using the current method in the selected nine proteins of Table II are presented in Table IV.

In all of the previous studies, the identification of the domain interface residues has been carried out separately from that of the identification of domains. The difference in the accessible surface area of residues when present in multi-domains and single domains is the factor considered in identifying the domain interface residues. In the current work, we have presented a simple and single-numeric computation to identify the structural domains as well as the domain interface residues from protein three-dimensional structures. The same computation gives both the domain definitions and the interface residues between domains. Furthermore, the concepts from graph theory have been elegantly incorporated into protein structures for the analysis of multi-domain proteins.

Limitations of the Present Method

In the above presentation, we have demonstrated the power of the graph spectral method. In this section, we summarize the critical assessment of the performance of this method. The present method identifies domains in protein structures based on the fact that the connections within the domain are higher than across the domain and this difference is reflected in the 2evc plots of PScG and PBG. Hence, the main limitation of this method is that it identifies only structural modules as domains and no functional aspect can be incorporated or obtained from it. For instance, any enzyme where the catalytic residues are contributed from two separate structural domains (like Pepsin, 5PEP) would be classified as a two-domained protein by the present method because of its structural modules, though independently both the domains may not have any biochemical or biological function. Such a problem is encountered in most of the automatic domain identification methods and is not unique to our method; only SCOP processes the domain assignment with functional factor built into it.

Next, if we consider cases of swapped domains, where a part of one chain is structurally swapped with the same of the other chain as in the case of 1DZ3 (Sporulation response regulator Spo0A), the present method assigns the swapped regions to the chain to which it is structurally closer rather than the one to which it sequentially belongs. The method depends on the cutoff that needs to be employed by the user to obtain the correct domain definitions, which is in turn dependent on the compactness and packing density of the protein. On the one hand it is advantageous to have a variable parameter that can be optimized for each protein and on the other hand, the user has to decide the right cutoff. However, we have provided a fair rule that can be followed by the user to choose the right cutoff, using radius of gyration as a measure of the packing density of the protein.

Conclusions

We have presented a novel method for the identification of domains in protein structures by a graph theoretic

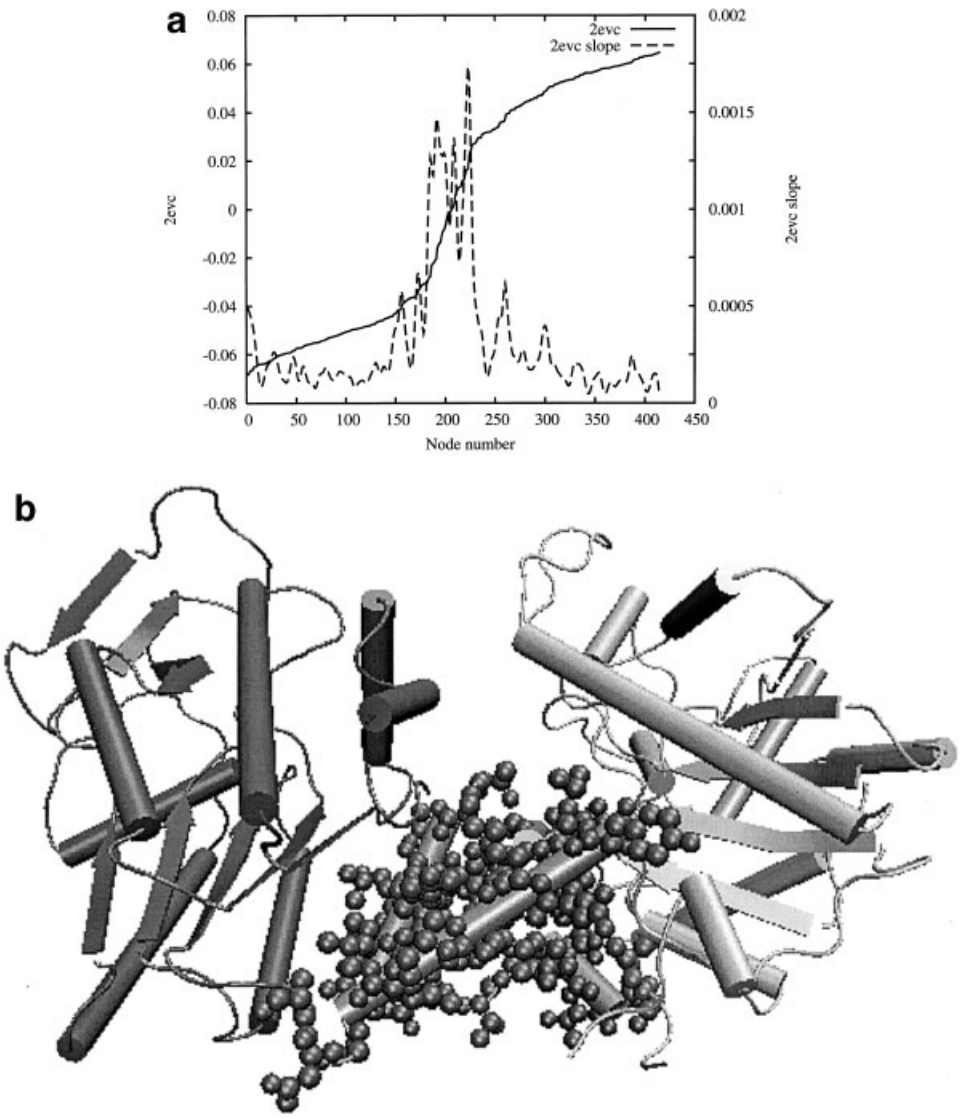

Fig. 6. **a**: The 2evc and 2evc smoothened slope plots for 16PK obtained from PScG. **b**: Domains and domain interface residues in phosphoglycerate kinase (16PK) obtained from the present method using PScG. The cartoon representation shows the two different domains in different shades of gray while the van der Waal's representation shows the interface residues.

technique. The concept is based on the representation of the protein structure topology in the form of a graph. Two types of graph representations are used. The protein backbone graph (PBG) is based on the C^{α} coordinates, which give the gross domain information and is a simple technique to use. On the other hand, the protein side-chain graph (PScG) representation is slightly more involved; the interaction among the side-chains of the amino acids residues are explicitly considered and quantified as the interaction strength. We find that an interaction cutoff of 4% is generally suitable for the identification of domains and domain interfaces in proteins. The optimal cutoff that needs to be employed for domain identification varies from protein to protein due to differences in the compactness of various proteins. We provide an easy method to identify the optimal cutoff for a protein using the radius of gyration as a measure of the compactness of the protein.

The residue–residue connectivity information is translated into a global matrix form called the Laplacian matrix. The solution to such a matrix carries the domain information. The vector components of the second lowest eigenvalue, which gives the clustering information can also give the information on the subclusters, which are identified as structural domains in proteins.

The advantage of this method is that an accurate domain definition can be obtained by a single numeric computation. Furthermore, the residues at the domain interfaces can also be obtained by the same computation. The graph spectral parameters obtained by this method can be easily incorporated into a protein coordinate file

TABLE IV. Interface Residues Obtained Using PScG

Pdb code	Interface residues from PScG
1YGE	20–23, 97–99, 106–112, 122–183, 242–246, 519–522, 642–644, 769–776
1YGE (12% Cutoff)	14–22, 104–140, 769–772, 195–198, 211–215, 248–251, 525–530, 546–548, 623–627, 650–654, 782–785, 264–267, 500–502, 602–612, 696, 697, 755, 756, 796–808, 824–827, 351–354, 406, 407, 428–434, 449–451, 476–479, 715–725
1LAM	68–74, 96–114, 140–154
1SMD	268–278, 399–415, 420–422, 484–486
3GRS	52–65, 104–113, 158–163, 175–179, 288–295, 338–341, 363–370, 449–454
5PEP	1–29, 50–52, 112–122
1SU4	124–128, 154–158, 213, 327, 328, 332–334, 341, 342, 345, 357–365, 386–390, 436, 440–444, 550–562, 598–602, 634, 638, 644, 645, 697, 725, 726, 729, 732, 733, 746, 747, 817
16PK	166–272, 192–203, 382–386, 392–418
1EPW	929–944, 964, 965, 985–989, 995–998, 1043–1049, 1076–1078, 1085–1087, 1127–1130, 1282–1284, 607–615, 636–642, 788, 829, 830, 855–859, 867, 868, 875–878, 910–919, 1028–1030, 437, 461, 462, 471, 472, 604, 622, 667, 668–676, 690, 694, 703, 707, 711, 767, 768, 810–817, 67, 68, 208–214, 245–247, 261–264, 268, 275, 279, 284, 286, 287, 370–372, 376, 430, 431, 479–481, 521, 522, 735–753
1AFB[a]	94–100, 225, 226 of A,B and C

[a]Trimer, composed of A, B and C chains.

(PDB) to automatically display the different domains of a protein in different colors. Additionally, we use a variable cutoff parameter that can be objectively chosen based on the compactness of the protein and also helps in identifying domains and interfaces at a desired level of resolution. This aids in correlating the discrepancies in domain definitions of different expert methods. Thus, we have a simple and elegant method based on graph theory for the identification of domains and domain interfaces in protein structures.

ACKNOWLEDGMENTS

R.K.S. and K.V.B. would like to thank the Department of Science and Technology, India and the Center for Scientific and Industrial Research, India respectively, for the fellowships. We acknowledge the Computational Genomics Initiative at the Indian Institute of Science, funded by the Department of Biotechnology (D.B.T.), India, for support.

REFERENCES

1. Richardson JS. The anatomy and taxonomy of protein structure. Adv Protein Chem 1981;34:167–339.
2. Swindells MB. A procedure for the automatic determination of hydrophobic cores in protein structures. Protein Sci 1995;4:93–102.
3. Holm L, Sander C. The FSSP database of structurally aligned protein fold families, Nucl Acid Res 1994;22:3600–3609.
4. Siddiqui AS, Barton GJ. Continous and dicontinous domains: an algorithm for the automatic generation of reliable protein domain definitions. Protein Sci 1995;4:872–884.
5. Islam SA, Luo J, Sternberg MJE. Identification and analysis of domains in proteins. Protein Eng 1995;8:513–525.
6. Sowdhamini R, Rufino SD, Blundell TL. A database of globular protein structural domains: clustering of representative family members into similar folds, Fold Des 1996;1:209–220.
7. Jones S, Stewart M, Michie A, Swindells MB, Orengo C, Thorton JM. Domain assignment for protein structures using a consensus approach: characterization and analysis. Protein Science 1998;7:233–242.
8. Conte LL, Ailey B, Hubbard TJP, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of protein database. Nucleic Acid Res 2000;28:257–259.
9. Veretnik S, Bourne PE, Alexandrov NN, Shindyalov IN. Toward consistent assignment of structural domains in proteins. J Mol Biol 2004;339:647–678.
10. Y. Xu, D. Xu, Gabow HN. Protein domain decomposition using a graph-theoretic approach. Bioinformatics 2000;16:1091–1104.
11. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
12. Patra S, Vishveshwara S. Backbone cluster identification in proteins by a graph theoretical method. Biophys Chem 2000;84:13–25.
13. Kannan N, Vishveshwara S. Identification of side-chain clusters in protein structures by a graph spectral method. J Mol Biol 1999;292:441–464.
14. Hall KM. An r-dimensional quadratic placement algorithm. Manage Sci 1970;17:219–229.
15. Creighton TE. Proteins: structures and molecular properties, 2nd Ed., New York, San Francisco: Freeman; 1996.
16. Humphrey W, Dalke A, Schulten K. VMD—Visual Molecular Dynamics. J Mol Graph 1996;14:33–38.
17. Jones S, Marin A,Thornton JM. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. Protein Eng 2000;13:77–82.