PRAMANA
— journal of
physics

© Indian Academy of Sciences

Vol. 71, No. 2
August 2008
pp. 369–378

Analysis of protein folds using protein contact networks

PANKAJ BARAH and SOMDATTA SINHA*
Mathematical Modelling and Computational Biology Group, Centre for Cellular and Molecular Biology (CSIR), Uppal Road, Hyderabad 500 007, India
*Corresponding author. E-mail: sinha@ccmb.res.in

Abstract. Proteins are important biomolecules, which perform diverse structural and functional roles in living systems. Starting from a linear chain of amino acids, proteins fold to different secondary structures, which then fold through short- and long-range interactions to give rise to the final three-dimensional shapes useful to carry out the biophysical and biochemical functions. Proteins are defined as having a common 'fold' if they have major secondary structural elements with same topological connections. It is known that folding mechanisms are largely determined by a protein's topology rather than its interatomic interactions. The native state protein structures can, thus, be modelled, using a graph-theoretical approach, as coarse-grained networks of amino acid residues as 'nodes' and the inter-residue interactions/contacts as 'links'. Using the network representation of protein structures and their 2D contact maps, we have identified the conserved contact patterns (groups of contacts) representing two typical folds – the EF-hand and the ubiquitin-like folds. Our results suggest that this direct and computationally simple methodology can be used to infer about the presence of specific folds from the protein's contact map alone.

Keywords. Protein structure; network; contact map; fold recognition; EF-hand; ubiquitin-like.

PACS Nos 89.75.-k; 87.14.Ee; 05.65.+b

1. Introduction

Proteins are linear polymers built from 20 different amino acids, which share a common structural feature – an alpha carbon to which an amino group, a carboxyl group and a variable side chain are bonded. The side chains of the amino acids have varying physicochemical properties thereby interacting with each other in different ways. Proteins are synthesized in the ribosome inside the cell as nascent polypeptide chains, which are not their functional forms. Due to extensive hydrogen bonding, this nascent polypeptide chain gives rise to different regularly repeating local secondary structural elements (alpha helices, beta strands). Short- and long-range non-local interactions among these secondary structural elements lead the protein to fold into a functionally active, native, three-dimensional, tertiary structure. Often more than one tertiary subunits interact among themselves to form quaternary structures. The interesting fact is that, even though the protein world is extremely diverged at the primary sequence level, the way they fold to produce the three-dimensional architecture are limited to only a few hundreds in nature. What makes diverse proteins fold to similar 3D topology is an open question.

Recently, the native structures of proteins have been studied using the network formalism as protein contact networks (PCN) by considering the three-dimensional structure of a protein as a network of amino acids [1–10]. Here each constituent amino acid is considered as a node and the contacts among the amino acids are the links/edges. Much of the work was centered on understanding the relation of the network properties (e.g., degree distribution, clustering, assortativity, closeness, etc.) to protein's structural features such as, stability, folding/unfolding, communication and interfaces. Unlike most other networks, PCNs have some specific features – it is constrained by its backbone (i.e., the chain of amino acids), and specific groups of short-range contacts can correspond to distinctive secondary structural features.

Classification of proteins is based on both structural and functional relatedness. According to the central dogma of bioinformatics, the primary amino acid sequence dictates structure, and the specific three-dimensional structure determines the function of a protein. The database 'structural classification of proteins' (SCOP) [11] is a well-recognized classification system of proteins, which is based on manual inspection of structural similarities in the three-dimensional topology from structural data. As defined by SCOP, there exist several hierarchies. The principal levels are family, superfamily, fold and class. According to SCOP, proteins clustered together into families are evolutionary-related where the percentage of sequence similarity is high. The proteins grouped into a common superfamily have probable common evolutionary origin, as they have low sequence identities, but their functional features show a common evolutionary origin. Proteins placed together in the same fold category may not have a common evolutionary origin. Proteins share a common fold if they have similar arrangement of major secondary structural elements (SSE) with the same topological connections. In spite of having low sequence similarity, all the proteins from the same fold may show close topological relatedness. Different proteins from the same fold often have the same peripheral SSEs and turn regions that differ in size and conformation. Based on the abundance of different secondary structures (alpha helices and beta sheets) folds are again grouped into four classes of proteins, namely all alpha, all beta, alpha + beta and alpha/beta.

Three-dimensional structures of complex macromolecules, such as proteins, can be represented well by its two-dimensional distance map and its contact approximation. Here contact is defined according to different distance thresholds (R_c). For proteins such maps can be defined at different resolutions, starting from finest atomistic details to the simple amino acid coarse contact level [12]. This level can even be extended up to the level of protein secondary structural elements, as seen in protein topology cartoons [13]. Even though these 2D coarse contact maps do not contain all the details of protein structures at the atomistic resolution, they are still quite useful tools for studying complex protein structures, which provide a good representation of the overall topology of a protein, and capture most of the relevant structural information. These contact maps can be useful for rapid comparison of protein structures as protein fingerprints [14]. It has also been shown [15] that knowing the contact positions of residues in the protein contact maps are extremely useful in protein structure and folding prediction. This has been well demonstrated in the 'critical assessment of techniques for protein structure prediction' (CASP3 and CASP4) [16,17].

In this work, using contact map formalism, we have used the coarse-grained network description of native protein structures to assess the possibility of developing a methodology to understand (a) specific features in the network representation that correspond to the distinct secondary structural elements of the proteins and (b) structural and contact features that may be conserved in the networks of diverse proteins that belong to the same fold. Taking three proteins each from two specific folds – ubiquitin-like and E-F hand-like – as examples, we show that different secondary structural elements can be identified in the contact maps, and the short- and long-range contacts among different secondary structural elements in each fold-type can be identified from the contact matrices. The visualization in the ring graph can easily correspond to the information in the 2D matrix. This methodology can be applied to other folds, and offers a simple procedure for detecting specific folds from experimental coordinate data.

2. Methodology

2.1 *Construction of PCN and contact map*

The native-state protein structures are modelled as networks made of their constituent amino acids and their non-covalent (within a threshold distance) and covalent (peptide bonds between consecutive amino acids) interactions. The coarse-grained PCN is generated from the structural data available in the protein data bank (PDB) [18,19], by considering only the C_α atom of each amino acid as a 'node' and any two amino acids are said to be in spatial contact if there exist a threshold distance R_c ($\leq$7 Å) between their C_α atoms. The threshold distance can be varied from a very high, fine-grained resolution to a very low, coarse-grained resolution. Here the choice of threshold distance was done based on the inter-residue chemical interactions [8].

This distance map is a 2D symmetric, square matrix where the entry (i,j) represents the distance between the nodes i and j along the protein primary sequence chain from the N to C terminal. The contact map of the PCN is a Boolean matrix (adjacency matrix) of pair-wise inter-residue contact representation based on the threshold distance. A contact map (M) for a protein with n_r residues is a matrix of order $n_r \times n_r$ whose elements are defined as $M_{ij} = 1$, if residues i and j are in contact, else $M_{ij} = 0$.

The protein 3D structures have been visualized using molecular graphics software Pymol [20], which is freely available. The adjacency matrices are visualized using MATLAB 6.1 [21]. PCN is visualized using the free software PAJEK [22]. In this study we have used circular representation (ring graph), where the nodes are plotted on a circle starting from the N to C terminal of the protein primary sequence chain. For the two-dimensional view, the Kamada–Kawai lay-out of the networks has

Table 1. Six proteins from two structural folds. The first three proteins (1BTOa, 1H8Ca, 1I35a) belong to ubiquitin-like fold and last three proteins (1C07a, 1FI6a, 1MHO) belong to E-F hand-like fold.

PDB ID	Name of protein	Source organism	Function	Size (n_r)	Resolution (if X-ray)
1BTOa	Ubiquitin-like protein, rub1 [23]	*A. thaliana*	Signalling protein	73	1.70
1H8Ca	Fas-associated factor 1 [24]	*Homo sapiens*	Apoptosis	82	–
1I35a	Protein kinase byr2 [25]	*S. pombe*	Transferase	95	1.50
1C07a	Epidermal growth factor receptor pathway substrate 15 [26]	*Homo sapiens*	Signalling protein	95	–
1FI6a	reps1 Eh domain protein [27]	*Mus musculus*	Ca^{2+} binding	92	–
1MHO-	S-100 protein [28]	*Bos taurus*	Ca^{2+} binding	88	2.0

been used in PAJEK. Protein topology cartoons have been obtained from PDBsum (http://www.ebi.ac.uk/pdbsum/).

2.2 *Data*

The fold information has been adopted from the SCOP. Table 1 lists six proteins selected from the Ubiquitin-like and E-F hand-like folds belonging to different functional classes and sources. Structural data (X-ray, NMR) are downloaded from the PDB. Figure 1 shows the 3D structures (ribbon diagram) and the topology cartoons of the two folds considered.

Ubiquitin-like fold: Ubiquitin's main role is in targeting proteins for degradation. This fold is characterized by a beta-grasp fold – beta-beta-alpha-beta-beta-alpha-beta [24]. As shown in figure 1a, if the beta strands are numbered according to their occurrence in the protein backbone from N terminus to the C terminus, in this beta-grasp fold, the five beta strands are arranged into a mixed sheet in the order 2-1-5-3-4. The longer first helix packs across the first three strands of the sheet, and a second shorter 3_{10} helix is located in an extended loop connecting strands 4 and 5.

E-F hand-like fold: This motif takes its name from the traditional nomenclature used in describing the protein parvalbumin [29], which contains three such motifs. The EF-hands can be divided into two classes: signalling proteins and buffering/transport proteins [30]. These proteins (except Calbindin d9k) typically undergo a calcium-dependent conformational change, which opens a target-binding site. As shown in figure 1b, the E-F hand is a helix-turn-helix structural motif in proteins. It consists of two alpha helices positioned roughly perpendicular to one another and linked by a short loop region (usually about 12 amino acids) that often binds calcium ions.

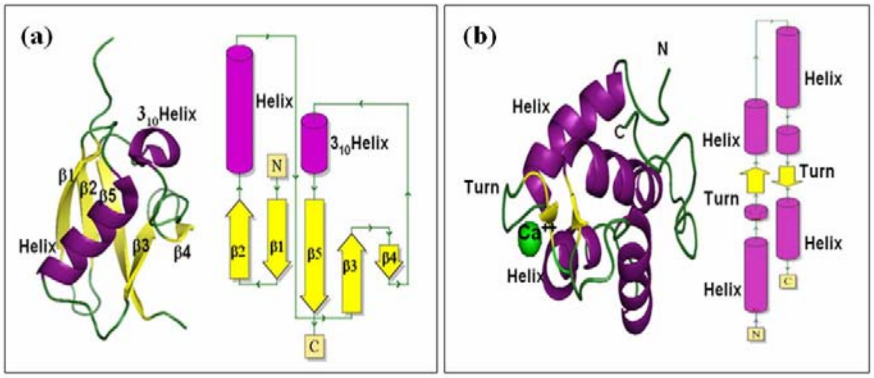

Figure 1. The ubiquitin-like and E-F hand-like folds. (**a**) Protein 1H8C from ubiquitin-like fold and its topology cartoon characterized by the presence of beta-grasp fold and (**b**) protein 1C07 from E-F hand-like fold and its topology cartoon characterized by the presence of helix-turn-helix motif.

3. Results

3.1 *Exploring protein contact maps*

Figure 2 presents different types of visualizations used in this work, with protein 2IGD as example (IGG-binding domain from *Streptococcus*) of size 61 amino acids. The topology cartoon shown in figure 2b gives a simplified view of the 3D structure.

In the 2D contact map (figure 2c), diagonal elements are the contacts along the protein backbone. Helices appear as thick bands along the main diagonal as every helical turn involves neighbouring intrahelical contacts between one amino acid and its four successors. The contact map is compared with the secondary structural elements (SSE) in the topology cartoon, which shows that the parallel and anti-parallel contacts between the SSEs are seen as cluster of points parallel and perpendicular to the main diagonal (or, the backbone) in the contact map. In the contact map, contacts between residues closer in the linear polypeptide chain are placed closer to the diagonal, and contacts between far (along the backbone) residues (long-range contacts) are placed away from the diagonal as is seen in the case of the parallel contacts between the first and last beta strands. In the ring graph shown in figure 2d, one can clearly see the long-range interactions that bring residues, placed far apart in primary structure, in close proximity. Here, the parallel contacts appear as twisted lines (as between the first and last beta strands), and anti-parallel contacts appear as parallel lines. Figure 2e shows the PCN of 2IGD as a two-dimensional graph. It is clear that a combination of contact map and ring graph representation of the PCN corresponding to a protein's native structure can give useful information about its contact patterns and the SSEs.

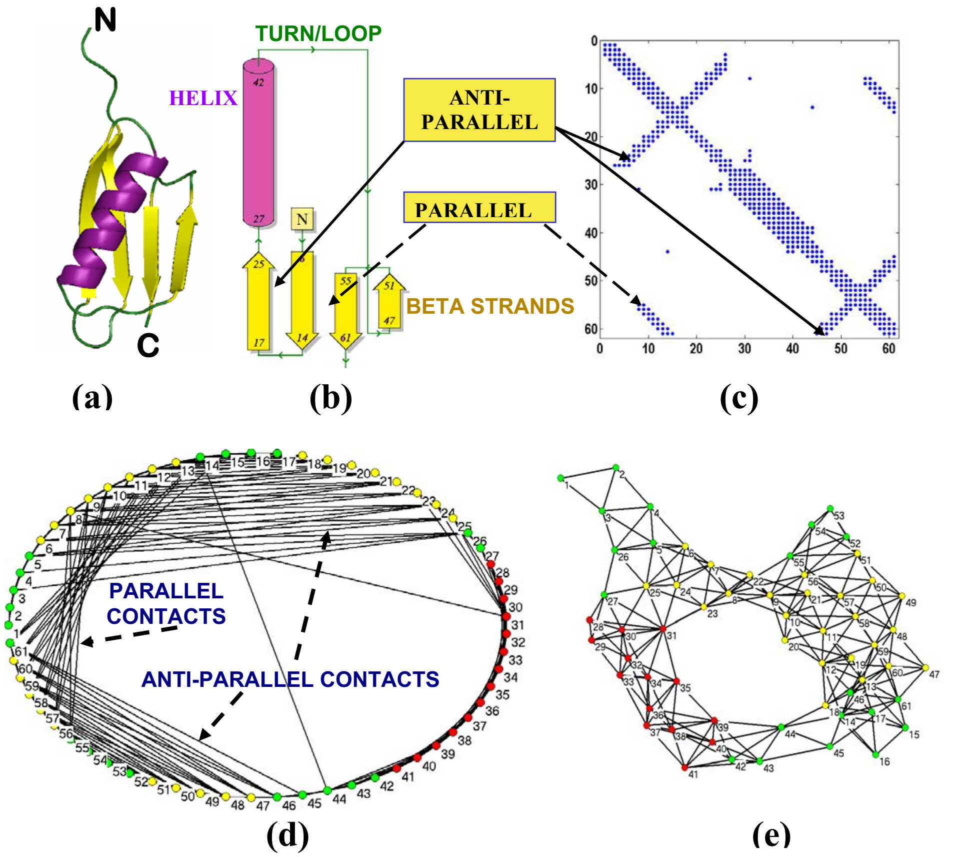

Figure 2. Different visualizations of protein 2IGD: (**a**) 3D structure, (**b**) topology cartoon, (**c**) 2D contact map, (**d**) ring graph and (**e**) 2D representation of the PCN.

3.2 *Identifying conserved contact patterns in ubiquitin-like and E-F hand-like folds*

In figures 3 and 4 we present the contact maps and the ring graphs of three proteins each from ubiquitin-like fold and E-F hand-like fold respectively. In both the figures, the contacts and SSEs are denoted in the contact map and ring graph of the first protein (figures 3a and 4a). Figures 3 shows the ubiquitin-like fold characterized by the presence of beta-grasp fold, which is a sequence of SSEs given by beta-beta-alpha-beta-beta-alpha-beta. The contact pattern of beta-grasp fold can be easily identified from the contact map, as well as, from the ring graph visualization of protein 1H8C (figure 3a). The typical contact patterns of this fold (beta-grasp fold) are the antiparallel contacts between 1st and 2nd beta strands, intrahelical contacts of helix 1, followed by anti-parallel contacts between two far apart beta turns, the intrahelical contacts of 3_{10} helix, and lastly, the parallel contacts between 1st and 5th beta strands. Figures 3b and 3c show that the contact patterns, which

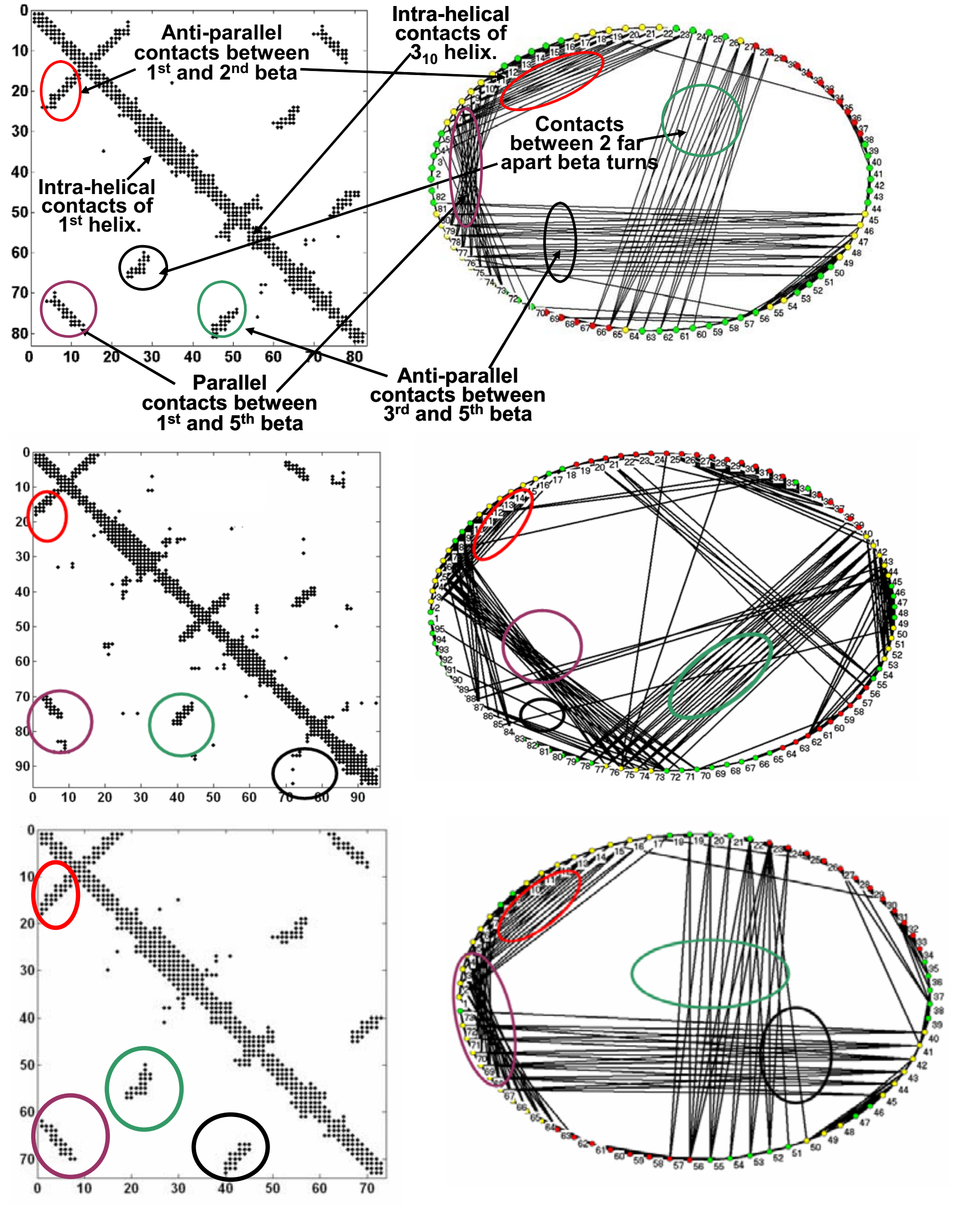
Anti-parallel contacts between 1st and 2nd beta
Intra-helical contacts of 3_{10} helix.
Contacts between 2 far apart beta turns
Intra-helical contacts of 1st helix.
Parallel contacts between 1st and 5th beta
Anti-parallel contacts between 3rd and 5th beta
0
10
20
30
40
50
60
70
80
90

Figure 3. Contact patterns of ubiquitin-like fold (beta-grasp fold) in the contact map and ring graph of (**a**) 1H8C, (**b**) 1I35 and (**c**) 1BTO.

we observed in 1H8C are also distinctly found in the other proteins (1I35 and 1BTO) having the same fold. Thus, these are conserved contact patterns for proteins having the ubiquitin-like fold. Similarly, figure 4 shows the three proteins having the E-F hand-like fold, which is characterized by the presence of the helix-turn-helix motif

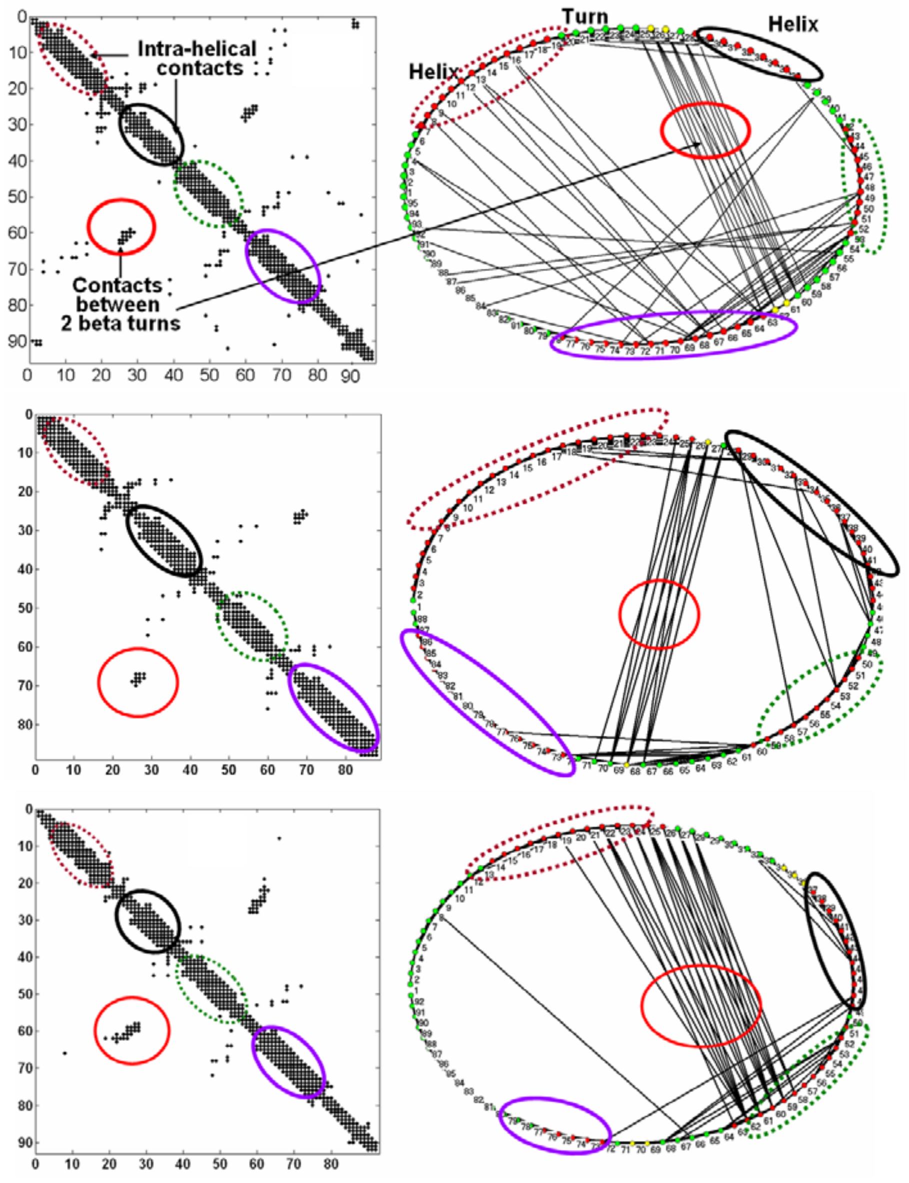

Figure 4. Contact patterns of EF hand-like fold (helix-turn-helix fold) in the contact map and ring graph of (**a**) 1CO7, (**b**) 1FI6 and (**c**) 1MH0.

that binds calcium ion (see figure 1). In the contact map and ring graph of protein 1C07 (figure 4a), are shown clearly the contact patterns for the helix-turn-helix. Similar circles represent the participating helices in the motif, and it is clear that there are two helix-turn-helix motifs in this protein. The conserved contact pattern for this fold can be identified from the intrahelical contacts of the 1st helix followed by the turn and the intrahelical contacts of the 2nd helix, which is present in the other two proteins (1FI6, 1MH0) (also shown in figures 4b and 4c).

4. Conclusion

We have attempted to decipher the arrangements of different secondary structural elements (folds) in the 3D structure of proteins available in PDB, from their coarse-grained protein contact networks, which do not consider the atomistic details. The 'contact map', represented compactly in a symmetrical, square, Boolean matrix of pair-wise, inter-residue contacts, of the 3D conformation of the protein, and the ring graph visualization of the PCN are found to be best suited for this purpose. Irrespective of the size and type of the proteins, we have identified the conserved contact patterns (groups of contacts) representing a typical fold – for the EF hand-like and ubiquitin-like folds. These contact patterns can be clearly identified in the contact map and ring graph even without the help of the visualization of the three-dimensional structure. Thus, this coarse-grained approach can be used to locate different types of folds or their combinations without an atomistic analysis of the protein structure from the experimental data. We believe that other motifs and folds can also be studied using this methodology.

Acknowledgements

PB thanks DBT, Govt. of India for financial support. PB gratefully acknowledges R R Sarkar, S Bhattacharyya, Ranjib Dey and N S Shiju Lal for valuable suggestions and help.

References

[1] L H Greene and V A Higman, *J. Mol. Biol.* **334**, 781 (2003)
[2] N Kannan and S Vishveshwara, *J. Mol. Biol.* **292**, 441 (1999)
[3] K V Brinda and S Vishveshwara, *Biophys. J.* **89**, 4159 (2005)
[4] K V Brinda and S Vishveshwara, *Biophys. J.* **92**, 2523 (2007)
[5] Md. Aftabuddina and S Kundu, *Physica* **A396**, 896 (2006)
[6] M Vendruscolo, N V Dokholyan, E Paci and M Karplus, *Phys. Rev.* **E65**, 061910 (2002)
[7] U K Muppirala and Zhijun Li, *Protein Engineering, Design & Selection* **19**, 265 (2006)
[8] G Bagler and Somdatta Sinha, *Physica* **A346**, 27 (2005)
[9] G Bagler and Somdatta Sinha, *Bioinformatics* **23**, 1760 (2007)
[10] N S Shiju Lal and Somdatta Sinha, *Proceedings of the 11th ADNAT Convention on Advances in Structural Biology and Structure Prediction* 134 (2007)
[11] Murzin *et al*, *J. Mol. Biol.* **247**, 536 (1995)
[12] G Pollastri and P Baldi, *Bioinformatics* **18**, S62 (2002)
[13] D R Westhead, D C Hatton and J M Thornton, *Trends in Biochem. Sci.* **23**, 35 (1998)
[14] A Godzik, J Skolnick and A Kolinski, *J. Mol. Biol.* **227**, 227 (1999)
[15] J Selbig and P Argos, *Proteins: Struct. Funct. Genet.* **31**, 172 (1998)

[16] A M Lesk, L L Conte and T J P Hubbard, *Proteins* **45(S5)**, 98 (2001)

[17] A R Ortiz, A Kolinski, P Rotkiewiez and J Skolnick, *Proteins Suppl.* **3**, 177 (1999)

[18] Bernstein *et al*, *J. Mol. Biol.* **112**, 535 (1977)

[19] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov and P E Bourne, *Nucleic Acids Res.* **28**, 235 (2000)

[20] W L DeLano, *The PyMOL Molecular Graphics System* (2002), www.pymol.org

[21] The MathWorks, Inc. (www.mathworks.com/)

[22] V Batagelj and A Mrvar, in *Graph drawing software* edited by M Jünger and P Mutzel (Springer, Berlin, 2003) p. 77

[23] C Rao-Naik, W delaCruz, J M Laplaza, S Tan, J Callis and A J Fisher, *J. Biol. Chem.* **273**, 34976 (1998)

[24] A Buchberger, M J Howard, M Proctor and M Bycroft, *J. Mol. Biol.* **307**, 17 (2001)

[25] W Gronwald, F Huber, P Grunewald, M Sporner, S Wohlgemuth, C Herrmann and H R Kalbitzer, *Structure* **9**, 1029 (2001)

[26] J L Enmon, T de Beer and M Overduin, *Biochemistry* **39**, 4309 (2000)

[27] S Kim, D N Cullis, L A Feig and J D Baleja, *Biochemistry* **40**, 6776 (2001)

[28] M Andersson, A Malmendal, S Linse, I Ivarsson, S Forsen and L A Svensson, *Protein Sci.* **6**, 1139 (1997)

[29] J P Declercq, C Evrard, V Lamzin and J Parello, *Protein Sci.* **8**, 2194 (1999)

[30] S Nakayama, N D Moncrief and R H Kretsinger, *J. Mol. Evol.* **34**, 416 (1992)