Analysis of core–periphery organization in protein contact networks reveals groups of structurally and functionally critical residues

ARNOLD EMERSON ISAAC[1] and SITABHRA SINHA[2,*]

[1]*Bioinformatics Division, School of Bio Sciences and Technology, VIT University, Vellore, India*

[2]*The Institute of Mathematical Sciences, Chennai, India*

**Corresponding author (Email, sitabhra@imsc.res.in)*

The representation of proteins as networks of interacting amino acids, referred to as protein contact networks (PCN), and their subsequent analyses using graph theoretic tools, can provide novel insights into the key functional roles of specific groups of residues. We have characterized the networks corresponding to the native states of 66 proteins (belonging to different families) in terms of their core–periphery organization. The resulting hierarchical classification of the amino acid constituents of a protein arranges the residues into successive layers – having higher *core order* – with increasing connection density, ranging from a sparsely linked *periphery* to a densely intra-connected *core* (distinct from the earlier concept of protein core defined in terms of the three-dimensional geometry of the native state, which has least solvent accessibility). Our results show that residues in the inner cores are more conserved than those at the periphery. Underlining the functional importance of the network core, we see that the receptor sites for known ligand molecules of most proteins occur in the innermost core. Furthermore, the association of residues with structural pockets and cavities in binding or active sites increases with the core order. From mutation sensitivity analysis, we show that the probability of deleterious or intolerant mutations also increases with the core order. We also show that stabilization centre residues are in the innermost cores, suggesting that the network core is critically important in maintaining the structural stability of the protein. A publicly available Web resource for performing core–periphery analysis of any protein whose native state is known has been made available by us at *http://www.imsc.res.in/~sitabhra/proteinKcore/index.html.*

[Isaac AE and Sinha S 2015 Analysis of core–periphery organization in protein contact networks reveals groups of structurally and functionally critical residues. *J. Biosci.* **40** 683–699] **DOI** 10.1007/s12038-015-9554-0

1. Introduction

Proteins, biological macromolecules that are essential for the structure, function and regulation of a living cell, are linear chains of amino acids that fold into three-dimensional structures comprising different secondary structural elements, such as helices, sheets and coils, by making short- and long-range contacts between amino acid residues along the chain. The overall shape of the protein (and very often, its function) is determined by its folded, tertiary native structure. From the point of view of its three-dimensional geometry, the core of a folded protein is the region which has least solvent accessibility, consisting of amino acids that are more hydrophobic than the residues on the exposed surface. Determining how the different regions of a protein contribute to its function and structural stability continues to be an important scientific challenge that has inspired the development of various methods for analysing the structure of proteins. A relatively recent and promising

Keywords. Core–periphery; hierarchical core decomposition; protein contact network

Supplementary materials pertaining to this article are available on the *Journal of Biosciences* Website at *http://www.ias.ac.in/jbiosci/oct2015/supp/Emerson.pdf*

Published online: 29 September 2015

approach to understanding the relative contributions of the different components of a protein has been inspired by advances in the study of complex networks that occur in different domains (Newman 2010). Representation of protein native state as a network of interacting amino acids (e.g. see Di Paola *et al.* 2012 for a review) has allowed the application of an array of graph theoretic techniques and concepts, such as that of small-world networks (Watts and Strogatz 1998), for analysing protein structure.

A network is defined by a set of nodes or vertices, and a set of links or edges that connect certain pairs of nodes. In a protein contact network (PCN), the nodes correspond to the amino acid residues that the protein consists of, while the links are defined by information about the physical adjacency of each pair of residues in the three-dimensional structure of the folded protein. In particular, a network is defined by stating that, when the distance between two amino acids (to calculate the distance, one could consider every atom that make up the amino acids, or only the C-α atoms that belong to the backbone of the protein) is less than a specified threshold or 'cut-off distance' value – usually corresponding to the range of significant non-covalent interaction between the amino acids, the pair are considered to be linked, else they are not connected. The result can be displayed in an adjacency matrix **A**, each of whose elements indicate whether the pair of residues corresponding to a row (i, say) and column (j, say) are connected ($A_{ij} = 1$) or not ($A_{ij} = 0$). When focusing on long-range interactions between residues that otherwise occur far from each other in the primary sequence of the protein, one could also define a lower threshold to neglect all connections between pairs of amino acids whose distance is less than this value. It is clear from the above that the choice of the cut-off distances largely determines the nature of interactions that are included in the analysis (Afonnikov *et al.* 2006; da Silveira *et al.* 2009). Most studies on PCNs have only considered an upper threshold cut-off, taken to be around 8Å (Brinda *et al.* 2005; Aftabuddin and Kundu 2006; Barah and Sinha 2008). A few studies have introduced a lower limit (around 4Å) to eliminate contacts arising from proximity in the primary sequence.

Protein contact networks comprising long- as well as short-range interactions have been shown to be small world, defined as the coexistence of short average path length and high clustering (Vendruscolo *et al.* 2002; Bagler and Sinha 2005), with the distribution of degree (i.e. the number of links that a node has) having a Poisson-like nature (Greene and Higman 2003). Other graph theoretic properties of the adjacency matrix for PCNs have been studied in detail (Vendruscolo *et al.* 2001; Vendruscolo *et al.* 2002; Vishveshwara *et al.* 2002; del Sol *et al.* 2006a, b; Bagler and Sinha 2007; Vishveshwara *et al.* 2009; Vijayabaskar and Vishveshwara 2010; Vendruscolo 2011). For instance, it has been shown that the shortest path lengths in the network and residue fluctuations are highly correlated (Atilgan *et al.* 2004). Correlation between the most interconnected residues at protein–protein interfaces and residues that contribute the most to the binding free energy has also been observed (del Sol and O'Meara 2005). In addition, study of a large set of enzymes has shown that active site residues tend to be highly central, suggesting that these positions are crucial for the transmission of information between the residues in the protein (Amitai *et al.* 2004). Other studies have considered PCNs as weighted graphs by associating variable strengths with the connections between pairs of residues, e.g. using the information about side chains (Brinda and Vishveshwara 2005; Brinda *et al.* 2005; Brinda *et al.* 2005; Kundu 2005; Aftabuddin and Kundu 2006; Vishveshwara *et al.* 2009; Brinda *et al.* 2010). The study of interactions in a complex system can often provide insights into the critical factors governing its stability (Jeong *et al.* 2001; del Sol *et al.* 2006a, b). Significant progress has been made in this direction over the past decade by studies considering energy fluctuations and their correlations, locations of conserved residues, stability of the native state, binding between proteins and between proteins and ligands from the perspective of network representation of a protein (Bahar *et al.* 1997; Haliloglu *et al.* 1997; Bahar *et al.* 1998; Demirel *et al.* 1998; Bahar *et al.* 1999; Vendruscolo *et al.* 2002; Amitai *et al.* 2004; Haliloglu *et al.* 2005; Ertekin *et al.* 2006; Haliloglu *et al.* 2008; Haliloglu and Erman 2009; Yogurtcu *et al.* 2009). From the point of view of applications, it is important to note that network-based studies of proteins can help in identifying drug target candidates (Csermely *et al.* 2013). Different measures of centrality have been investigated in PCNs in order to understand the role played by different residues in the architectural organization of a protein (Hu *et al.* 2014). Several studies have also explored the occurrence of intermediate-scale (or mesoscopic) features of PCNs such as modularity by exploring sub-domain architecture using community detection algorithms (Hleap *et al.* 2013). For instance, the small copper protein *azurin* has recently been decomposed into modules of strongly interacting residues that highlight different structural and functional features (Tasdighian *et al.* 2013).

In this study we analysed PCNs by focusing on another prominent mesoscopic organizational feature of many complex networks, viz. the existence of a densely intra-connected core and a relatively sparsely connected periphery (Everett and Borgatti 1999). Initially introduced in the context of social networks, such core–periphery organization has later been reported in a large number of biological networks (Holme 2005; Wuchty and Almaas 2005). The original concept of a two-class division of nodes into core and periphery has now been generalized to embrace networks having a hierarchical arrangements of layers (defining progressively higher-order orders) characterized by the intra-connectivity within members of the inner cores becoming more dense. A decomposition technique for revealing this hierarchical core–periphery structure through recursive pruning of nodes based on their degree (Seidman 1983) has been

successfully applied on a number of real-world networks, including the internet (Alvarez-Hamelin *et al.* 2005; Carmi *et al*. 2007), the neuronal network of the nematode *Caenorhabditis elegans* (Chatterjee and Sinha 2007) and the protein interaction network of *Escherichia coli* (Lin *et al.* 2009). Recently, this decomposition technique has been used to disentangle the hierarchical structure of Internet router-level connection topology (Zhang *et al.* 2009), to show that software systems are organized in a defined hierarchy of increasing centrality from outside to inside (Zhang *et al.* 2010) and to demonstrate that, in disease spreading models, the most efficient spreaders are those located within the core of the network (Kitsak *et al.* 2010).

Here we have applied this network decomposition technique to protein contact networks in order to demonstrate their core–periphery hierarchical organization and see whether residues belonging to the innermost cores play a critical role in the function or structural stability of the protein. Note that our use of the term core in the context of PCNs is distinct from its earlier usage, defined in terms of the three-dimensional geometry of the native state, as the part that has least solvent accessibility (see, for example, Toth-Petroczy and Tawfik 2011). We see that residues in the inner cores indeed appear to be evolutionarily more conserved than those belonging to the lower order (outer) cores, i.e. the periphery. We also see that the receptor sites for known ligand molecules of most proteins occur in the innermost core which underlines the functional importance of the network core. Furthermore, the association of residues with structural pockets and cavities in binding or active sites increases with the core order. From mutation sensitivity analysis, we show that the probability of deleterious or intolerant mutations also increases with the core order. We also show that stabilization centre residues are in the innermost cores, suggesting that the network core is critically important in maintaining the structural stability of the protein. A publicly available Web resource has been made available by us for performing core–periphery analysis of any protein whose native state is known at *http://www.imsc.res.in/~sitabhra/proteinKcore/index.html*.

2. Materials and methods

2.1 *Protein structure analysis*

We compiled a set of 66 non-redundant protein structures obtained from the Protein Data Bank (*http://www.rcsb.org/pdb*) that include 41 from distinct protein families, including both enzymes and non-enzymes, and 25 belonging to different pathogenic organisms. Supplementary table 1 provides details of each protein in the dataset with protein name and PDB id.

2.2 *Protein contact network*

The three-dimensional structure of proteins is modeled as an undirected network, with each node of the network corresponding to an amino acid of the protein. The edges of the network represent the interaction between the amino acids, which are determined as follows. We calculate the minimum Euclidean distance in three dimensions between any two residues i and j as $d(i,j) = \min_{\alpha,\beta}[(x^{\alpha}_i - x^{\beta}_j)^2 + (y^{\alpha}_i - y^{\beta}_j)^2 + (z^{\alpha}_i - z^{\beta}_j)^2]^{1/2}$, where the labels $\alpha=1,\ldots,n_i$ and $\beta=1,\ldots,n_j$ run over all the atoms in the two amino acids. Thus, the distance between two residues is calculated by finding the minimum of all their pair-wise inter-atomic distances. A pair of amino acids are said to be connected in the corresponding contact network if their distance $d(i,j)$ is less than a threshold value., i.e. $PCN(i,j)=1$ if $d(i,j)<d_{cutoff}$ and otherwise $PCN(i,j)=0$. For most of our analysis we have chosen this cut-off for considering two residues to be connected as d_{cutoff} = 5Å, which approximates the upper limit for attractive London–van der Waals interactions (del Sol et al. 2006a, b). We have explicitly verified that small variations in d_{cutoff} do not affect our results significantly. Note that, in the literature, protein contact networks have been constructed using several different criteria (e.g. see Di Paola *et al.* 2012 for a review). Networks constructed by using only the distance between C-α atoms of a pair of amino acids have typically used threshold values between 8–10Å, while those obtained by considering distance between all atoms of the corresponding amino acids have used threshold values around 5Å.

2.3 *Long-range interaction network*

The long-range interaction network (LIN) is constructed from the PCN by removing links to sequential neighbours (i.e. residues that lie in adjacent positions along the protein sequence). First, the cumulative distance matrix (CDM=$\{d^{cum}\}$) between all pairs of amino acids is constructed. To do this for a protein having N residues, we initially calculate the minimum Euclidean distance $d(i,i+1)$, $i=1,\ldots,N-1$, between every pair of neighbouring residues in the protein sequence. Then, the cumulative distance between any two residues is obtained by summing the nearest neighbour distances between all residue pairs that lie between i and j along the protein sequence, i.e. $d^{cum}(i,j) = d(i,i+1)+d(i+1,i+2)+\ldots+d(i+m,j)$ if there are m residues separating i and j along the protein sequence. The sequence adjacency matrix (SAM) is then constructed from the CDM as follows: if the cumulative distance along the sequence between any pair of residues i and j is less than d_{cutoff} (=5Å in our analysis), then $SAM(i,j)=1$, otherwise $SAM(i,j)=0$. Thus, a pair of nodes which are connected in the network represented by SAM are neighbours along the sequence. A connection is considered to be part of the LIN if it belongs to

Table 1. Folding nucleus residues of proteins that belong to the inner core

PDB ID and chain identifier	Protein name	Folding nucleus residue not present in inner core	Inner core residues*
1NLO - (c)	Sh3 domain	**50**	**10** 11 12 13 14 15 **16** 23 **24** 25 **26** 27 28 29 **30** 31 **32** 33 **34** 42 43 44 **45** 46 **47** 54 55 **56** 57 58 60 **61** 62
1YPC - (I)	Chymotrypsin inhibitor 2	-	21 22 23 24 25 26 27 28 29 30 31 32 33 34 **35** 36 37 38 **39** 40 41 42 43 44 46 47 48 49 50 51 53 64 65 **66** 67 **68** 69 70 71 74 75 **76** 77 78 79 80 81 82
1RIS - (A)	Ribosomal protein s6	-	1 2 3 4 5 **6** 7 **8** 9 **10** 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 **26** 27 28 29 **30** 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 48 55 57 58 59 **60** 61 62 **63** 64 **65** 66 **67** 68 69 71 72 73 74 **75** 76 77 78 **79** 80 82 84 85 86 87 88 89 90 91 92 93
1URN - (A)	Protein (U1A)	-	11 **12** 13 **14** 15 **17** **26** 27 **30** 31 32 33 **34** 35 36 37 38 **40** 42 **43** 44 **45** 54 **55** 56 57 **58** 59 64 65 66 67 68 **69** 70 71 72 73 82 83 **84** 85 86
2AIT - (A)	Tendamistat	**11,59**	5 6 7 8 9 **12** 13 14 15 **16** 20 21 **22** 23 24 **25** 31 32 **33** **34** **35** 36 37 42 43 44 45 46 47 48 52 53 54 55 **56** 57 58 67 69 **70** **71** 72
1APS - (A)	Acylphosphatase	**45**	6 7 8 9 10 **11** 12 **13** 22 23 24 **25** 26 27 28 29 **30** 31 32 33 34 **35** 36 37 38 39 40 **47** 48 49 50 51 52 53 **54** 55 56 57 58 59 60 61 62 63 64 65 66 67 68 75 77 78 79 80 81 82 83 84 89 91 92 93 **94** 95 96 98

*Residues marked in bold belong to the folding nucleus.

PCN but not to the network represented by SAM. In other words, LIN(i,j)=1 if PCN(i,j)=1 and SAM(i,j)=0. For all other cases, LIN(i,j)=0. This allows us to consider only connections between positional neighbours who are not sequential neighbours, i.e. residues that are otherwise far apart in the protein sequence which come close to each other only as a result of folding of the three-dimensional protein structure.

2.4 *Core decomposition*

2.4.1 *k-Core decomposition:* To identify the core–periphery organization of the protein contact network and long-range interaction network they were subjected to core decomposition. The k-core of a network is defined as a subnetwork which exclusively contains nodes which connect to at least k other nodes in the same subnetwork (Seidman 1983). The cores of different orders of a network can be obtained by *iteratively* removing all nodes which have less than k connections with other residues (k = 1, 2, …). This is done by first identifying all nodes whose degree (i.e. number of connections) is less than k. After removing these, the network is re-analysed to determine if the removal of these nodes has resulted in other nodes (which originally had degree > k) having now less than k connections. If such nodes are identified, then they are removed, and the process is continued, until no more nodes can be removed. The resulting subnetwork is called the k-core of the network.

2.4.2 *Randomization of k-core:* To determine the statistical significance of the properties calculated for members of an empirically determined k-core, we compared them to the mean and variance of the corresponding values obtained for a randomized ensemble. Each randomized k-core in the ensemble is obtained by random selection without replacement of N_k residues from the protein, where N_k is the size of the empirically determined k-core. The randomized ensemble for every protein considered was generated by constructing 100 such randomized k-cores.

A possible alternative would have been to randomize the network keeping its degree conserved and shuffling the nodes. However, in the context of a protein this is unphysical as it would result in corresponding protein structures that may be impossible to generate in three-dimensional physical space.

2.5 *Core analysis*

2.5.1 *Solvent Accessibility:* Accessible surface area (ASA) or the solvent accessibility of amino acids in a protein is defined as the relative surface area that can be in contact with a solvent. ASA for each amino acid is calculated using the

ASAView server (*http://www.netasa.org/*). The obtained values for all the amino acids in a protein are divided into 3 classes: least accessible (ASA=0–20%), moderately accessible (ASA=20–50%) and highly accessible (ASA>50%) with respect to the solvent (Shander *et al.* 2004).

2.5.2 *Determination of centre of mass of a protein:* To determine the relation between the network core and structural core of a protein, we determined the physical location of the *k*-core residues relative to the centre of mass of the protein whose position coordinates (x_{CM}, y_{CM}, z_{CM}) were calculated as

$$x_{CM} = \frac{\sum_{i=1}^{N} m_i x_i}{\sum_{i=1}^{N} m_i}, \quad y_{CM} = \frac{\sum_{i=1}^{N} m_i y_i}{\sum_{i=1}^{N} m_i}, \quad z_{CM} = \frac{\sum_{i=1}^{N} m_i z_i}{\sum_{i=1}^{N} m_i},$$

where (x_i,y_i,z_i) are the cartesian coordinates of the *i*-th atom and m_i is its atomic mass. The distance of a particular residue *j* (whose position coordinates are assumed to be same as that of the C-α atom within it) from the protein centre of mass is given by: $d_{CM}(j) = \sqrt{\left((x_j - x_{CM})^2 + (y_j - y_{CM})^2 + (z_j - z_{CM})^2\right)}$, and the average of this quantity for all residues in a *k*-core provides the mean position of the core measured from the centre of mass.

2.6 *Functional importance of residues*

2.6.1 *Conservation of protein residues:* The conservation score for each amino acid residue in a protein is obtained via the *Consurf* server (*http://consurf.tau.ac.il*) which is a relative measure of the evolutionary conservation at each sequence site of the target protein with the *lowest* score representing the *most conserved* position. It uses ClustalW Multiple Sequence Alignment for calculating the scores of all residues and then performs a normalization to make the mean score =0 with standard deviation = 1. The continuous conservation scores are partitioned into a discrete scale of 9 bins for visualization, such that bin 9 contains the most conserved positions and bin 1 contains the most variable positions. We have used a BLAST cut-off value of 0.0001 in order to minimize the possibility of erroneously including non-homologous sequences. To increase the accuracy of the conservation scores, the phylogenetic tree has been also taken into account which helps in better identification of the amino-acid replacements or substitution that could have occurred in the family of homologous sequences through evolution (Glaser *et al.* 2003; Landau *et al.* 2005; Ashkenazy *et al.* 2010).

2.6.2 *Role of residues in ligand–protein interactions:* The functional importance of a residue in a protein may be a result of it taking part in ligand-protein interaction. This information is obtained from the LPC CSU server (*http://bip.weizmann.ac.il/oca-bin/lpccsu*) which determines the contacting residues in a protein with a ligand and the type of interactions they undergo (e.g. hydrophobic–hydrophobic, aromatic–aromatic, etc.) based on a detailed analysis of the inter-atomic contacts and interface complementarity (Sobolev *et al.* 1999). For one particular protein, for which information could not be obtained from LPC CSU, the PDBSum database was used (*http://www.ebi.ac.uk/pdbsum/*).

2.6.3 *Association of residues with structural pockets and cavities in protein:* Functionally important sites in a protein are often associated with structural pockets and cavities. We identified residues associated with pockets/cavities by using the CASTp server (*http://sts.bioengr.uic.edu/castp/*) which utilizes the weighted Delauney triangulation and the alpha complex for shape measurements. The area and volume of each pocket and its cavities are measured, both in the solvent accessible surface (SA, Richards' surface) and the molecular surface (MS, Connolly's surface). Other dimensions such as that of mouth openings, area of the openings, circumference of mouth lips, in both SA and MS surfaces for each pocket are also obtained using the same method (Dundas *et al.* 2006).

2.6.4 *Mutation sensitivity of residues:* Replacing a functionally important residue by any other residue (compared to other positions in the protein sequence) is more likely to affect the fitness of a protein and is, therefore, likely to be a deleterious mutation. We identify such sites that are intolerant to mutations using the SIFT server (*http://sift.jcvi.org/*) that predicts whether the substitution of a particular amino acid affects the function of the protein based on sequence homology and the physical properties of a residue (Ng and Henikoff 2003).We calculate the probability that replacing the *i*-th residue by any other amino acid results in impairment of protein function as $P_i = X/19$, where X is the number of amino acid replacements (out of the total of 19 possible) that is predicted by SIFT to be a deleterious mutation. By adding together this probability for all residues belonging to a core of a particular order k gives us the mutation sensitivity of the *k*-th core as:

$$P_k = \sum_{i=1}^{N_k} \frac{P_i}{N_k},$$

where N_k is the total number of residues in the *k*-core. This represents the mean probability that replacing any member of the *k*-core by another amino acid will result in seriously affecting the function of the protein.

2.6.5 *Identifying stabilizing residues:* Residues that presumably play an important role in stabilizing a protein can be potentially identified by combining the information about several of its attributes, such as, large surrounding hydrophobicity, high long-range order and conservation score and its membership in a stabilization centre (Magyar *et al.* 2005). We identify such sites using the SRide server (*http:// sride.enzim.hu/*) with the default conditions (*i*) surrounding hydrophobicity, i.e. the sum of hydrophobic indices of surrounding residues whose C-α atom are within a distance of 8Å of the C-α of the residue under consideration, H_P ≥20 kcal/mol, (ii) long-range order measured by the fraction of long-range contacts of the residue, LRO≥0.02, (iii) the residue belongs to a stabilization centre identified using the SCide server (*http://www.enzim.hu/scide*) and (iv) conservation score ≥6. The stabilizing residues (SRs) identified using this method typically corresponds to a small percentage of all the residues in the protein.

2.7 *Input and output data of the k-core server*

The input of the *k*-core server is the atomic coordinate file of the protein to be analysed. It can be specified by providing the four-letter PDB code. Alternatively, it can be any other atomic coordinate file in PDB format uploaded directly by the user. The *K*-core decomposition is carried out on the selected protein chain, the node type for constructing the protein contact network and cut-off values ranges from 5 to 12 which represent the intensity of the London–van der Waals interactions. The output of the server is a list of the cores with atomic coordinate files. The *k*-core server is located at *http://www.imsc.res.in/~sitabhra/proteinKcore/index.html*.

3. Results

3.1 *The importance of the protein core*

As already mentioned earlier, important residues in a protein have been sought to be identified using contact network properties such as degree and betweenness centrality (del Sol *et al.* 2006a, b). The core order of residues that we use as a distinguishing feature in this study has a crucial difference with these other measures used earlier. While degree and betweenness centrality are properties that are defined with respect to individual nodes, core order is measured with respect to a group of connected nodes. It is thus not a microscopic, i.e. node specific property but rather a mesoscopic feature of the contact network. Using this measure, nodes are not considered to be important merely on their own but rather because of the cluster to which they belong – in other words, it is the group as a whole which is identified as being important rather than its constituent members. While, the degree specifies residues that are in contact with many other residues and the betweenness centrality is used to find residues that act as *bridges* between different regions of the protein, the inner core helps distinguish strongly bound groups of residues that can function as a coherent unit. Figure 1 shows schematically the distinction between these measures in a situation where different sets of nodes are identified by the different properties used. In general, of course, a node can have high degree and/or high betweenness centrality, and also belong to the innermost core.

In this study, we identified the inner core residues of a large variety of different proteins and shown that these amino acids are functionally important, as shown by a variety of different measures, including, conservation, resistance to mutation, ligand interaction, etc.

The three-dimensional structural information about a protein is first used to construct the corresponding interaction network. Depending on whether we consider the interactions among residues that are positional neighbours as well as neighbours along the primary sequence or exclusively consider the former class of interactions, we define the protein contact network (PCN) and the long-range interaction network (LIN), respectively (for details, see the section on Methods). To obtain the cores of different orders in either of these networks we carry out the *k*-core decomposition technique (see the section on Methods). This procedure provides us with a nested hierarchy of protein network cores comprising a subset of residues of the protein that are increasingly inter-connected as the order increases. Not surprisingly, with increasing core order the number of residues belonging to that core steadily decreases for both the PCN and the LIN (figure 2A–B). For most proteins analysed in our study the innermost core for PCN ranged between orders of 6 and 8 while for LIN it ranged between 4 and 5 (supplementary table 1). The inset of figure 2 shows the relative size of the innermost core relative to the protein. While in this particular case, the inner core constitutes a substantial fraction of the entire set of residues belonging to the protein, there are situations where the inner core residues comprise a relatively small part of the entire protein (supplementary figure 1).

3.2 *The network core corresponds to the structural nucleus of a protein*

Before proceeding to understand the functional significance of the residues belonging to the innermost cores, we first seek to clarify their relation to the physical structure of the protein. In particular, we inquire as to whether the residues in the network periphery belong to the surface of the native conformation of the protein, and the network core

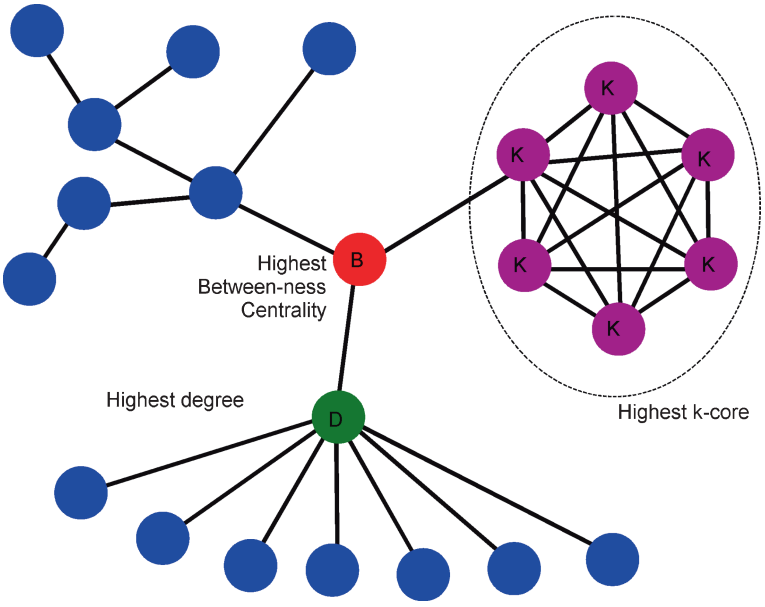

Figure 1. The importance of core order. Schematic diagram of a network indicating the distinction between nodes having highest degree centrality, highest betweenness centrality and highest core order. Note that, while the three properties are shown to belong to different nodes in this example, in general, there may exist a situation in which nodes may exhibit two or more of these properties in common.

corresponds to the structural nucleus. We approached this question by analysing the solvent accessibility of the residues in each core. Residues that are exposed on the outer surface of a folded protein are classified as having high accessibility, while those which are buried in the interior are less accessible. In between these two extremes are residues that are labeled as moderately accessible because they are only partially buried in the interior of the protein. Our analysis shows that the percentage of less accessible residues in a core increases with its order, while that of residues having high accessibility decreases (the percentage of moderately accessible residues show a slight decrease with increasing core order) (figure 3 and supplementary table 2). This suggests that the inner core of a protein has a significantly high representation from residues that also belong to the structural core.

As already mentioned in the beginning, our definition of a network core is distinct from the earlier usage of protein 'core' defined in terms of least solvent accessibility. However, we note that the groups of residues obtained from using these two different concepts do overlap. For the PCN, the average hydrophobic values are found to increase as the core order increases. This is true in 74 % of PCN and 81% for LIN (supplementary table 3). The average values have been calculated by summing all the hydrophobic index values for hydrophobic residues and divided by the total number of hydrophobic residues (respectively for neutral and hydrophilic residues).

We verify the conclusion concerning the strong correlation between network topological core and structural cores by calculating the average distance of the residues in each core from the centre of mass of the protein. Our calculations show that the residues belonging to the inner core are, in general, closer to the centre of mass of the protein than the residues belonging to the periphery (supplementary table 4). Our finding resonates with a recent study that evaluated proteins by measuring the distance of the surface residues from the protein centre of mass and has shown that, on average, the binding site residues are closer to the centre of mass than the non-binding surface residues (Nicola and Vakser 2007).

In addition to solvent accessibility, we have also considered whether the core decomposition method can help in identifying residues in the folding nucleus. When a protein folds or unfolds, it passes through many half-folded microstates, only a few of which can accumulate and be seen experimentally. The transition state (TS) is located in between the unfolding states and the native state on the free energy landscape (Abkevich *et al.* 1994; Fersht 1997; Pande *et al.* 1998). It has been observed that around the TS there are key contacts which are defined as folding nuclei (FNs), and the related residues of these contacts are known as folding nucleus residues (FNRs). Thus, the FNs and related FNRs play an essential role in the folding dynamics. We selected six protein structures (as shown in table 1) for which the FNRs had been identified experimentally, as well as, using simulation methods (Fersht 1995; Itzhaki *et al.* 1995;

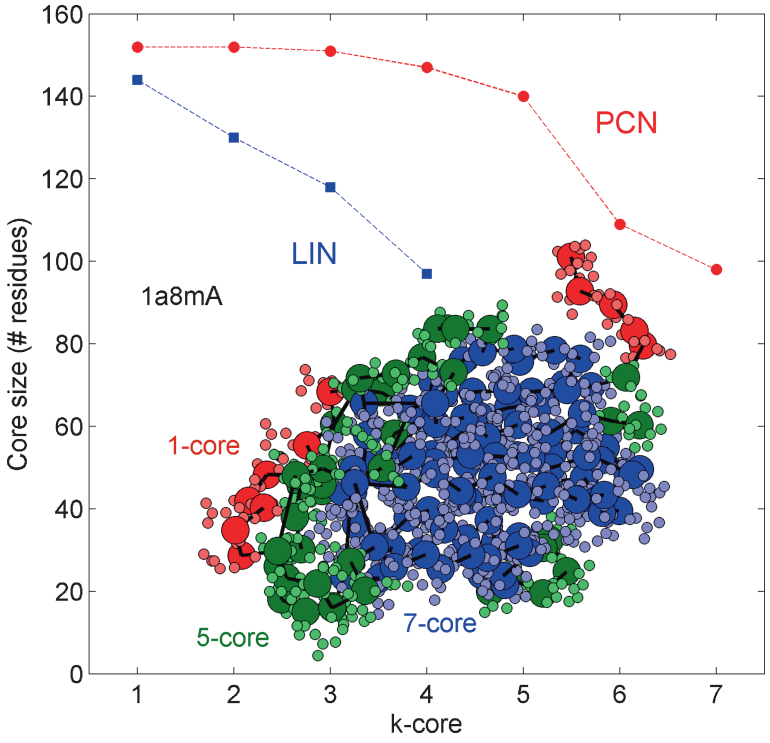

Figure 2. Sizes of the different cores for the interaction network of a protein.The number of residues in core of different order k for the (A) PCN and (B) LIN of the chain A (polymer 1) of the tumour necrosis factor (TNF) protein [PDB id: 1a8mA]. The inset shows a 3-Dimensional representation of the atomic composition of the protein, with the C-α atoms indicated as larger circles than the other atoms; the protein sequence 'backbone' is shown by connecting the neighbouring C-α atoms using solid lines. Residues in the innermost core (k=7) are indicated in blue while those belonging to 5-core but not in the 7-core are shown in green. Residues in the outer periphery (i.e. belonging to 1-core but not to 5 or higher order cores) are shown in red.

Grantcharova *et al.* 1998; Gruebele and Wolynes 1998; Riddle *et al.* 1999; Ternstrom *et al.* 1999; Clementi *et al.* 2000; Dokholyan *et al.* 2000; Li and Shakhnovich 2001; Vendruscolo *et al.* 2002; Hubner *et al.* 2004; Shen *et al.* 2005; Qin *et al.* 2006; Li *et al.* 2008). We found that the majority (if not all) of the FNRs are present in the inner core of each protein structure (table 1). This suggests that the k-core decomposition of a protein contact network can be used to predict the folding nucleus residues, which correlate strongly with the actual FNRs of the proteins used as examples here.

3.3 *Residues in inner cores are more conserved than those at the periphery*

To understand the significance of the residues belonging to the different cores, we initially analyse their degree of conservation (that measures its rate of evolutionary change) as a function of the core order. As the changes in different positions in a protein are not homogeneous but rather differ significantly, with some residues mutating rapidly (called 'variable' positions) relative to others (termed as 'conserved' positions), we pursue to determine if residues belonging to the inner cores are more conserved than those belonging to the periphery.

As mentioned in the Methods section, the conservation score for each residue is a relative measure of its evolutionary conservation, normalized so as to have zero mean and unit variance. Lower scores correspond to more conserved positions. As shown in figure 4, for both the PCN and the LIN, the relatively highly conserved residues are more numerous in the innermost residues. Instead of looking individually at the scores for each residue, we can instead perform an average of these scores for all members belonging to each core that would indicate whether the residues in the inner cores are more conserved in general (supplementary table 5). As shown in figure 5A, the average conservation score for constituent residues for each core decreases as the order increases, indicating that the innermost cores are

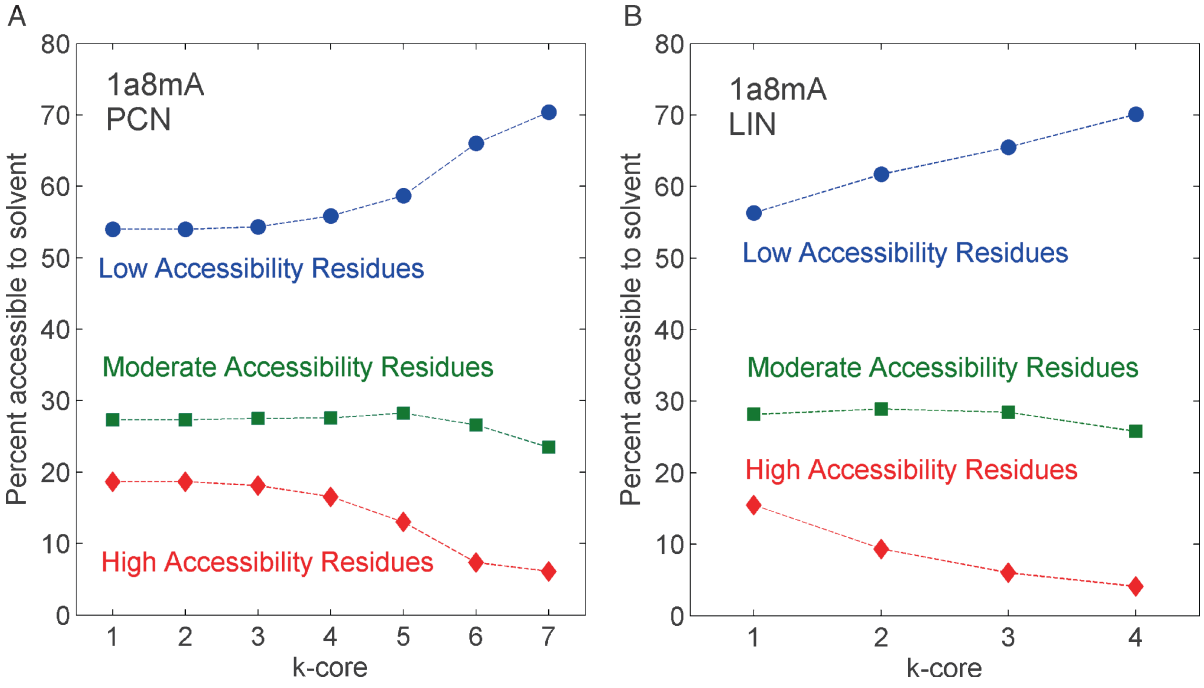

Figure 3. Residues of the protein belonging to inner core are much less accessible to solvent than those in the periphery. The fraction of highly solvent accessible (triangles), moderately solvent accessible (square) and less solvent accessible (circles) residues (measured in terms of percentage) in cores of different order for the (A) PCN and (B) LIN of chain A (polymer 1) of the tumour necrosis factor (TNF) protein [PDB id: 1a8mA]. With increasing core order the fraction of less accessible residues increase while that of more accessible residues decrease indicating the relative in-accessibility of the inner core to the solvent.

made up of highly conserved residues as compared to the outer periphery.

To verify whether the higher proportion of conserved residues that we observe in the inner cores is statistically significant, we compare the empirical values against the average conservation score for randomized cores of the same order. As shown in figure 5A, the average scores corresponding to the random cores do not show any significant deviation with core order, unlike the case for the actual protein. The deviation of the empirical data with core order is much greater than the error bars obtained from the random ensemble, suggesting that the highly conserved nature of the inner core residues is significant. The cumulative distribution of normalized conservation scores for the individual residues in the innermost core of a protein (figure 5B) also shows significant deviation from the corresponding distribution obtained from an ensemble of randomized *k*-cores having the same order. We have performed ANOVA test for statistical significance of the conservation scores of the innermost core residues at 95% confidence interval. Supplementary figure 2 shows that more than 85% of the randomized trials have *p*-value less than 0.05 confirming the statistically significant nature of the result. Supplementary table 6 compares between empirical conservation scores of core residues and the corresponding values for randomly selected residues to indicate the significance of the former.

Out of the 66 proteins that we had considered in our study only 6 proteins did not exhibit the trend of residues in the innermost core being more conserved: these are HIV-1 Protease (1a30), Annexin XII hexamer (1aei), Dihydrofolate Reductase (1aoe), Delta 2 Crystallin (1auw), Phosphate Regulon Transcriptional Regulatory Protein PHOB (1b00) and Acyl-CoA dehydrogenase (2Dvl). For all other proteins, we observed significant increase in the proportion of conserved residues as one progresses to the innermost core, which can thus be taken to be a general feature of proteins.

As evolutionary conservation of a residue may often be related to its functional importance for the protein, the above results strongly suggest that the inner cores contain a higher proportion of functionally critical sites. This leads us to further questions about what could be the possible functional roles of the inner core residues. With the aim of clarifying this, we analysed the nature of ligand interactions that each residue belonging to a core may be involved in.

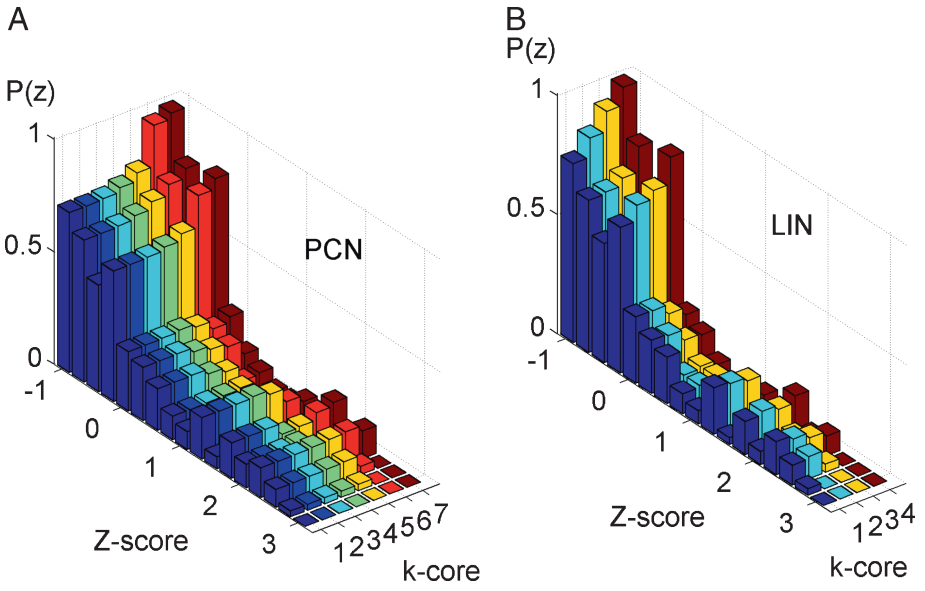

Figure 4. Inner-core residues of a protein are more likely to be conserved. The distribution of conservation z-scores for residues belonging to cores of different orders in the (A) PCN and (B) LIN of chain A (polymer 1) of the tumour necrosis factor (TNF) protein [PDB id: 1a8mA]. Residues having lower z-scores are more conserved; those belonging to the inner cores in both PCN and LIN show higher peaks at lower values of the z-score compared to residues in the outer cores indicating that the former are more likely to be conserved.

3.4 *Binding sites for ligand interactions are more likely to belong to the innermost core*

The function of many proteins is intimately related to their binding with specific substances (termed as 'ligand' molecules) to form a complex. The binding of the molecule to the receptor site of a protein alters the protein's chemical conformation that may initiate a specific biological action. To relate the conserved nature of a residue to its importance for the functioning of the protein, we first consider whether the residues that are part of the receptor sites of known ligand molecules belong to the innermost core of the protein.

As shown in table 2, residues belonging to the binding sites for known ligand molecules for most proteins are indeed observed to lie in the innermost core of the PCN. Figure 6 shows the interaction of a protein with its ligand molecule that clearly indicates that all the residues involved in the interaction belong to the innermost core (supplementary figure 3 shows yet another example, where the innermost core contributes the bulk if not all the residues interacting with the ligands – which in this case comprises a molecule as well as two carbon atoms). As ligand interactions is one of the most important functions that a protein is involved in, we can consider the above observation as validating our hypothesis that the more conserved nature of the inner core residues is related to their functional importance. Table 3 shows one of the few proteins which do not exhibit higher conservation for the inner core residues; as we can see, this may possibly be related to the fact that the residues of this protein which are involved in ligand binding do not belong exclusively to the higher order cores, but some of the receptor site residues can also belong to the periphery of the PCN that often correspond structurally to the surface of the protein molecule. Supplementary table 7 lists the percentage of residues in each core interacting with ligand molecules for all the proteins considered in our study which indicates that inner core residues have a far higher likelihood of belonging to a ligand binding site.

3.5 *Predicted binding and active sites in proteins are associated with inner core residues*

As specific ligand molecules (and their corresponding receptor sites in the protein) have not yet been identified for all the proteins that are being considered here, we have also considered possible binding and active sites that are often associated with structural pockets and cavities in the protein. Pockets are vacant regions having a concave geometry on the surface of the protein with an opening

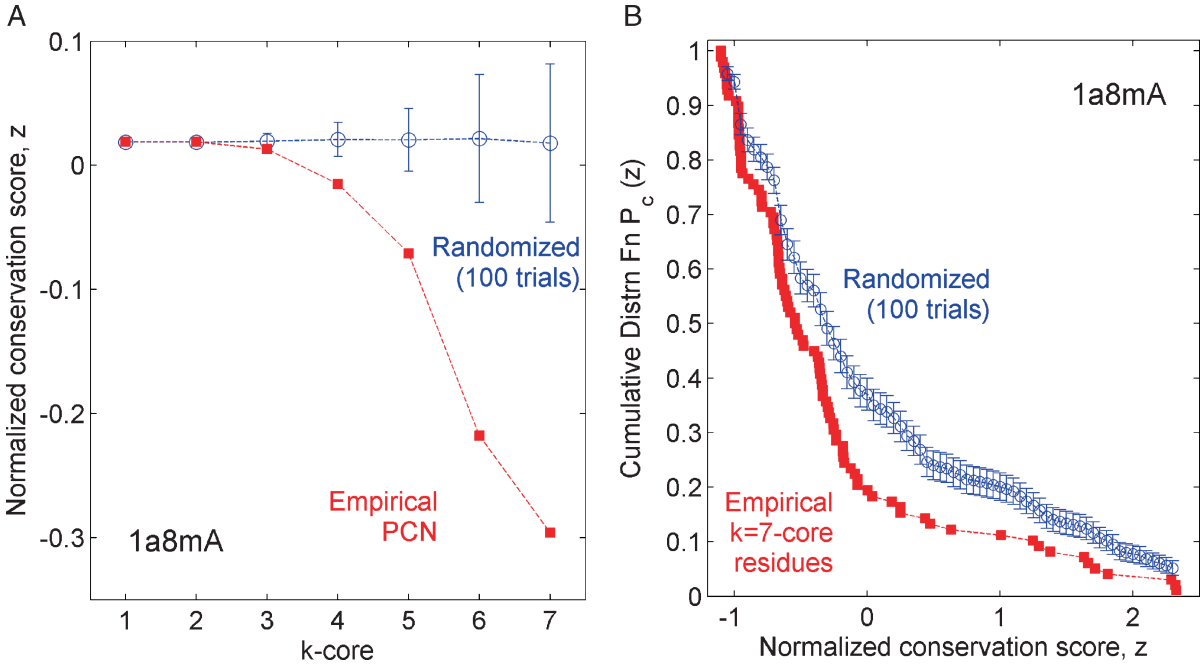

Figure 5. The conservation of the inner core of a protein is statistically significant. (A) The average normalized conservation z-score for all residues belonging to the different cores of the PCN for the chain A (polymer 1) of the tumour necrosis factor (TNF) protein [PDB id: 1a8mA]. For comparison we show the average normalized z-score for residues in cores of the same size constructed by randomly selecting residues from the protein. The result of averaging over 100 such randomizations are shown. (B) The cumulative probability distribution of the normalized conservation z-scores of the residues in the innermost core of the TNF protein. The corresponding randomized distribution is also shown which is calculated by averaging over an ensemble of 100 trials, each trial corresponding to constructing a set of z-scores of randomly selected residues belonging to the protein. The error bars represent the standard deviation over the 100 trials.

that connects their interior to the region exterior to the protein (Dundas *et al.* 2006). On the other hand, cavities are empty spaces in the interior of a protein that are inaccessible from the outside. We have obtained the identity of the residues belonging to all such surface accessible pockets, as well as, interior inaccessible cavities for the proteins included in our study. Analysis of the core-order membership of these residues (supplementary table 8) indicates that the fraction of residues which are a part of pockets/cavities, and hence, which are potentially part of binding or active sites, increases with the core order (83% of the proteins when considering PCN and 80% when we considered LIN).

Taken together, the strong correlation between the occurrence of a residue in higher order (i.e. inner) core and its likelihood of being part of a ligand interaction receptor region or a binding/active site, suggests one possible reason for the high degree of conservation of inner core residues as being due to their functional importance for the protein. However, binding of molecules is not the only important role that a specific set of residues in a protein may have. Residues may also be critical for the protein if they affect its structural stability. We can infer the importance of a residue for the viability of a protein by considering the consequences of mutating it.

3.6 *Mutations in inner core residues have a higher probability of being deleterious*

By replacing the actual amino acid occurring at a specific position in the primary sequence of a protein by any of the 19 other possibilities and verifying whether such a mutation is deleterious, we can obtain a quantitative measure of the critical importance of the residue for the protein. For instance, if the original residue is replaced by another amino acid and this does not correspond to a deleterious mutation, then we may conclude that the residue is not critical to the overall structural stability of the protein. On the other hand, if replacing the original residue by another amino acid *always* corresponds to a deleterious mutation, then it is reasonable to infer that the residue is extremely critical for the

Table 2. Innermost core residues of dehydroquinate synthase protein (1DQS) involved in interaction with ligands

k-Core	No of members	No of ligand interacting AA	% of AA	Amino acid no.
1	381	33	100	44,46,47,50,51,79,80,81,84,114,115,116,117,119,139,140,142,146, 147,149,152,161,162,179,182,183,184,187,190,194,250,286,287
2	381	33	100	44,46,47,50,51,79,80,81,84,114,115,116,117,119,139,140,142,146, 147,149,152,161,162,179,182,183,184,187,190,194,250,286,287
3	381	33	100	44,46,47,50,51,79,80,81,84,114,115,116,117,119,139,140,142,146, 147,149,152,161,162,179,182,183,184,187,190,194,250,286,287
4	381	33	100	44,46,47,50,51,79,80,81,84,114,115,116,117,119,139,140,142,146, 147,149,152,161,162,179,182,183,184,187,190,194,250,286,287
5	371	33	100	44,46,47,50,51,79,80,81,84,114,115,116,117,119,139,140,142,146, 147,149,152,161,162,179,182,183,184,187,190,194,250,286,287
6	357	33	100	44,46,47,50,51,79,80,81,84,114,115,116,117,119,139,140,142,146, 147,149,152,161,162,179,182,183,184,187,190,194,250,286,287
7	320	33	100	44,46,47,50,51,79,80,81,84,114,115,116,117,119,139,140,142,146, 147,149,152,161,162,179,182,183,184,187,190,194,250,286,287

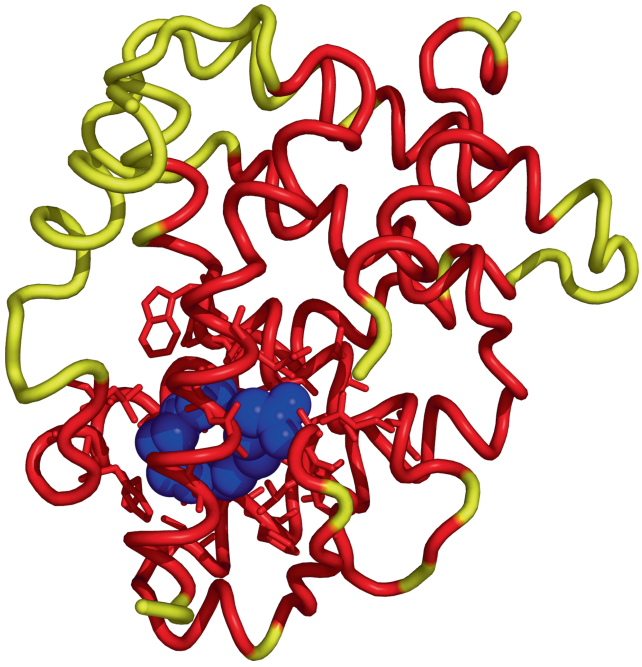

Figure 6. Residues in the inner cores are more likely to interact with an associated ligand. Cartoon tube representation of the tertiary structure of retinoid x receptor-alpha (PDB Id: 1dkfA) protein showing binding with a ligand molecule OLA (oleic acid, shown in blue). All the residues interacting with the ligand, indicated in stick format, belong to the innermost core (shown in red). The figure has been generated using the Open-Source PyMOL Molecular Graphic System, Version 0.99rc6.

overall stability of the protein. We can explore whether the residues belonging to the inner cores are more vital for the viability of the protein compared to those belonging to the outer cores or the periphery.

To check whether an amino acid substitution in the core affects the protein, we have replaced each position of a protein with the 19 possible substitutions of the original amino acid. This allows us to calculate the probability that a random mutation of the specific residue will be deleterious. We can then compare this probability for all residues belonging to the outer periphery with that for residues belonging to the inner core. Our results (supplementary table 9) show that the probability of deleterious or intolerant mutations tends to increase with PCN core order for 79% of the 61 proteins we considered while the probability remained essentially unchanged for a further 11%. For LIN, 77% of the 61 proteins showed an increase in the probability of deleterious or intolerant mutations while for 9% it remained the same, as core order was increased. Thus, we conclude that even when the core residues may not be directly involved in critical protein functions such as binding, they may be otherwise important in terms of ensuring the viability of a protein.

3.7 *Inner core residues impart structural stability*

The above measure is only an indirect indicator of the possible important role of a residue in ensuring the stability of a protein structure. To obtain a more direct criterion about how a specific residue stabilizes a protein we use a recently proposed identification procedure for stabilization centre residues (see the section on Methods for details) (Magyar *et al.* 2005). These residues appear to stabilize a protein

Table 3. Only five residues of HIV-1 protease protein (1a30) that interact with the ligand GLU-ASP-LEU belong to the innermost core

k-Core	No. of members	No. of ligand interacting AA	% of AA	Amino acid no.
1	99	8	100	25,28,29,30,47,48,49,50
2	99	8	100	25,28,29,30,47,48,49,50
3	95	8	100	25,28,29,30,47,48,49,50
4	93	8	100	25,28,29,30,47,48,49,50
5	78	5	62.5	25,28,29,30,47
6	68	5	62.5	25,28,29,30,47

structure through long-range interactions with their spatial, rather than sequential, neighbours. Our results (supplementary table 10) show that of the proteins considered in our study, 85% of the PCNs and 64% of the LINs have their stabilization centre residues in the innermost core. As stabilizing residues are also characterized by high degree of evolutionary conservation (Magyar *et al.* 2005), it reinforces our earlier observation that amino acids belonging to the inner cores are more conserved than those at the periphery. This is because these residues play an important role in imparting structural stability to the molecule, quite apart from their possible role as binding or active sites in the protein.

3.8 *Core analysis on the Web*

We have developed a Web server for performing *k*-core decomposition of proteins. It takes as input the three-dimensional structure of the protein which can be given either by simply writing the PDB ID (e.g, '1A3N', in which case the server directly takes the coordinates from the RCSB Protein Data Bank at *http://www.rcsb.org/*) or users can upload their own file containing the atomic coordinates. Once given this input, the user has to choose (i) which chain of the protein to analyse, (ii) the type of node to be used for constructing the contact network (i.e. whether to focus only on C-α atoms or whether all atom-atom interactions are to be considered) and (iii) the threshold d_{cutoff} for inter-atomic distance below which two atoms are assumed to be interacting (the user has the option to choose a value between 5 and 12 Angstroms). Given this information the server will generate a contact network and will perform *k*-core decomposition on it. As output the user can download files containing the residue id and atomic co-ordinates of the atoms belonging to the cores of different orders. It is also possible to visualize the different core structures using MDL Chime plug-in for the Web browser. The analysis results are freely available at *http://www.imsc.res.in/~sitabhra/proteinKcore/index.html*.

4. Discussion

Amino acids play a central role in biology both structurally, being the building blocks of proteins, as well as functionally, being the critical intermediaries of vital biochemical reactions, such as those which govern metabolism. Indeed, the principal information content of the genome is primarily concerned with specifying the composition of amino acids and the specific sequences in which to arrange them to construct all the proteins necessary for life. The protein sequence contains the necessary information that determines how it folds into a three-dimensional structure which is stable even in the highly noisy intra-cellular environment. The folding of proteins and their stability have been the subject of extensive research for decades but the many exciting questions in this field remain unresolved to date. Different approaches relying on structural features have been proposed to identify active sites in various proteins (Lichtarge *et al.* 1996; Aloy *et al.* 2001; Landgraf *et al.* 2001; Ondrechen *et al.* 2001). The representation of protein structures as interacting networks facilitates the analysis of topological characteristics, which could provide information about functionally important amino acids (Greene and Higman 2003).

In this study, we sought to understand whether functionally important residues are evolutionarily conserved and, moreover, whether the conserved residues within the protein core have an important role in maintaining the tertiary folded structure of the protein. To identify such critical residues we used core–periphery decomposition of the protein contact networks, as well as the corresponding long-range interaction networks. A set of 66 different structures spanning a broad range of protein families have been subjected to this analysis. A general feature observed in both the PCN and the LIN is that the size of the innermost core (i.e. the number of nodes comprising it) can differ substantially from that of the entire protein. We have examined the relation between the network core of a protein and its structural core by focusing on the solvent accessibility of the residues comprising each core order. We observe that the

percentage of less accessible residues increase in the innermost cores, implying that the core and periphery of the contact network has a correspondence to the structural core and surface of the protein, respectively.

Next, we verified the importance of the core residues by examining, e.g. the relative degree of conservation among the residues in different core orders. Of the 66 protein structures we analysed, 60 PCNs (i.e. 90%) show that the percentage of residues is highly conserved in the innermost core. However, we do note that there are a few exceptions, viz. proteins with PDB ids 1a30, 1aei, 1aoe, 1auw, 1b00 and 2dvl. We explicitly verified that the ligand interaction sites in these proteins occur at the periphery, which explains the relatively lower degree of conservation for the core residues. When we focus on the long-range interactions, we find that 89% of the LINs show high degree of conservation among the innermost core residues. For the exceptions, viz., proteins with PDB ids 1a30, 1aac, 1auw, 1cbr, 1vl4, 3e5y and 3js3, we verified that those proteins among this group (i.e. 1a30, 1aac, 1cbr, 3js3) which are known to bind with ligands have the ligand binding sites at the periphery. It is possible that the high degree of conservation of the inner core arises from the high inter-connectivity of its constituent elements. The role of the inner core in providing structural stability to the protein has been suggested by the observation that 57 PCNs and 58 LINs (of the 66 structures examined) have a very high probability of mutations in the innermost core being deleterious.

We explicitly checked that our results are not sensitively dependent on the specific value of the threshold distance d_{cutoff} used for defining the range of interaction between residues that is used in constructing the contact network. We also verified that defining the network in terms of distance measured between any pair of atoms or concentrating exclusively on the C-α atoms give similar results. Figure 7 shows the variation in size of cores as a function of their order for different definitions of distance and values of the threshold. As expected, using a higher d_{cutoff} results in a higher order of the innermost core as inclusion of many additional links makes the resulting network denser. Thus, the decomposition generates more layers before one arrives at the inner core. For similar reasons, considering only Cα atoms implies that relatively fewer number of links occur in the contact network; this, in turn, implies a lower order for the innermost core compared to the situation when all interatomic distances are considered.

The high likelihood of the innermost core of a protein hosting a ligand-binding site may have practical consequences in drug design. For instance, instead of considering the entire molecule, one can focus on the inner cores during the search for candidate sites in which a drug molecule can attach to a protein. This can significantly reduce the number of possibilities to be considered, thereby increasing the efficiency of the search procedure. The identification of core residues can also have potential significance in understanding the folding dynamics of a protein as it converges to its tertiary structure. It has been suggested that folding is initiated by the formation of a folding core which is also the final structure to break during denaturation (Haspel *et al.* 2003; Li 2009). Recent studies indicate that such folding cores have low solvent accessibility and high centrality (that implies they have a tightly packed network neighbourhood) (Li and Haiyan 2009). As the network core residues identified here have both of these properties, it is strongly suggestive of their possibly important role in coordinating the folding dynamics. We also suggest that concentrating on sequences

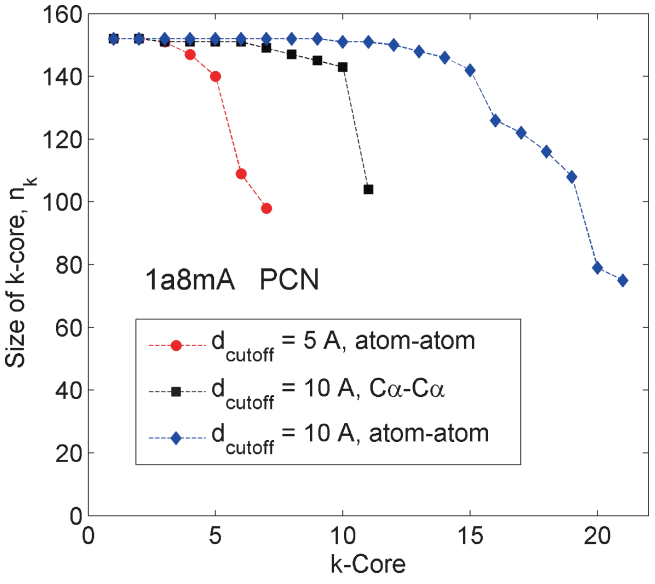

Figure 7. Robustness of observed core–periphery organization with respect to different methods of protein interaction network construction. Comparison of core–periphery organization of the PCN for chain A (polymer 1) of the tumour necrosis factor (TNF) protein [PDB id: 1a8mA] by using different methods of defining distance between residues and different thresholds d_{cutoff} for the distance between residues to define links in the PCN. In all the results described in the paper, the distance between residues *i* and *j* is measured by taking into account the Euclidean path length between coordinates of any atom in *i* with any atom in *j* (curve in red circles), and distances lower than d_{cutoff} = 5Å have been used to define the existence of a link in the PCN. Different core decompositions are obtained for the same protein if we use a different cut-off distance for defining the adjacency matrix, e.g. d_{cutoff}= 10Å (curve shown with blue diamonds) and different definition of distance between two residues, for instance, considering only the distance between the respective C-α atoms (curve shown with black squares). The qualitative nature of the curves, with core size decreasing with the order *k*, is similar in all cases.

comprising exclusively of the core residues during sequence alignment can be a more efficient method, e.g. during construction of phylogenetic trees, as it is precisely these segments which are the most conserved.

Acknowledgements

We would like to thank Indrani Bose and Somdatta Sinha for helpful discussions. We also thank the VIT University and IMSc for providing computational facilities.

References

Abkevich VI, Gutin AM and Shakhnovich EI 1994 Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33** 10026–10036

Afonnikov DA, Morozov AV and Kolchanov NA 2006 Prediction of contact numbers of amino acid residues using a neural network regression algorithm. *Biophysics* **51** 56–60

Aftabuddin M and Kundu S 2006 Weighted and unweighted network of amino acids within protein. *Phys. A.* **369** 895–904

Aloy P, Querol E, Aviles FX and Sternberg MJE 2001 Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **311** 395–408

Alvarez-Hamelin JI, Dall'Asta L, Barrat A and Vespignani A 2005 *k*-core decomposition: a tool for the visualization of large scale networks. arXiv preprint cs/0504107.

Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I and Pietrokovski S 2004 Network analysis of protein structures identifies functional residues. *J. Mol. Biol.* **344** 1135–1146

Ashkenazy H, Erez E, Martz E, Pupko T and Ben-Tal N 2010 ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res*. gkq399

Atilgan AR, Akan P and Baysal C 2004 Small-world communication of residues and significance for protein dynamics. *Biophys. J.* **86** 85–91

Bagler G and Sinha S 2005 Network properties of protein structures. *Phys. A.* **346** 27–33

Bagler G and Sinha S 2007 Assortative mixing in protein contact networks and protein folding kinetics. *Bioinformatics* **23** 1760–1767

Bahar I, Atilgan AR and Erman B 1997 Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* **2** 173–181

Bahar I, Atilgan AR, Demirel MC and Erman B 1998 Vibrational dynamics of folded proteins: significance of slow and fast motions in relation to function and stability. *Phys. Rev. Lett.* **80** 2733

Bahar I, Erman B, Jernigan RL, Atilgan AR and Covell DG 1999 Collective motions in HIV-1 reverse transcriptase: examination of flexibility and enzyme function. *J. Mol. Biol.* **285** 1023–1037

Barah P and Sinha S 2008 Analysis of protein folds using protein contact networks. *Pramana* **71** 369–378

Brinda K and Vishveshwara S 2005 A network representation of protein structures: implications for protein stability. *Biophys. J.* **89** 4159–4170

Brinda K, Surolia A and Vishveshwara S 2005 Insights into the quaternary association of proteins through structure graphs: a case study of lectins. *Biochem. J.* **391** 1–15

Brinda K, Vishveshwara S and Vishveshwara S 2010 Random network behaviour of protein structures. *Mol. Biosyst.* **6** 391–398

Carmi S, Havlin S, Kirkpatrick S, Shavitt Y and Shir E 2007 A model of Internet topology using *k*-shell decomposition. *Proc. Natl. Acad. Sci. USA* **104** 11150–11154

Chatterjee N and Sinha S 2007 Understanding the mind of a worm: hierarchical network structure underlying nervous system function in *C. elegans Prog. Brain Res.* **168** 145–153

Clementi C, Nymeyer H and Onuchic JN 2000 Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **298** 937–953

Csermely P, Korcsmaros T, Kiss HJM, London G and Nussinov R 2013 Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Ther.* **138** 333–408

da Silveira CH, Pires DEV, Minardi RC, Ribeiro C, Veloso CJM, Lopes JCD, Meira W, Neshich G, *et al*. 2009 Protein cutoff scanning: a comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins: Struct. Funct. Bioinf.* **74** 727–743

del Sol A and O'Meara P 2005 Small-world network approach to identify key residues in protein-protein interaction. *Proteins: Struct. Funct. Bioinf.* **58** 672–682

del Sol A, Fujihashi H, Amoros D and Nussinov R 2006a Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. *Protein Sci.* **15** 2120–2128

del Sol A, Fujihashi H, Amoros D and Nussinov R 2006b Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Mol. Syst. Biol.* **2**

Demirel MC, Atilgan AR, Bahar I, Jernigan RL and Erman B 1998 Identification of kinetically hot residues in proteins. *Protein Sci.* **7** 2522–2532

Di Paola L, De Ruvo M, Paci P, Santoni D and Giuliani A 2012 Protein contact networks: an emerging paradigm in chemistry. *Chem. Rev.* **113** 1598–1613

Dokholyan NV, Buldyrev SV, Stanley HE and Shakhnovich EI 2000 Identifying the protein folding nucleus using molecular dynamics. *J. Mol. Biol.* **296** 1183–1188

Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y and Liang J 2006 CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.* **34** W116–W118

Ertekin A, Nussinov R and Haliloglu T 2006 Association of putative concave protein-binding sites with the fluctuation behavior of residues. *Protein Sci.* **15** 2265–2277

Everett MG and Borgatti SP 1999 The centrality of groups and classes. *J. Math. Sociol*. **23** 181–201

Fersht AR 1995 Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc. Natl. Acad. Sci. USA* **92** 10869–10873

Fersht AR 1997 Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7** 3–9

Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E and Ben-Tal N 2003 ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **19** 163–164

Grantcharova VP, Riddle DS, Santiago JV and Baker D 1998 Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nat. Struct. Mol. Biol.* **5** 714–720

Greene LH and Higman VA 2003 Uncovering network systems within protein structures. *J. Mol. Biol.* **334** 781–791

Gruebele M and Wolynes PG 1998 Satisfying turns in folding transitions. *Nature Struct. Biol.* **5** 662–665

Haliloglu T and Erman B 2009 Analysis of correlations between energy and residue fluctuations in native proteins and determination of specific sites for binding. *Phys. Rev. Lett.* **102** 088103

Haliloglu T, Bahar I and Erman B 1997 Gaussian dynamics of folded proteins. *Phys. Rev. Lett.* **79** 3090

Haliloglu T, Keskin O, Ma B and Nussinov R 2005 How similar are protein folding and protein binding nuclei? Examination of vibrational motions of energy hot spots and conserved residues. *Biophys. J.* **88** 1552–1559

Haliloglu T, Seyrek E and Erman B 2008 Prediction of binding sites in receptor-ligand complexes with the Gaussian Network Model. *Phys. Rev. Lett.* **100** 228102

Haspel N, Tsai CJ, Wolfson H and Nussinov R 2003 Reducing the computational complexity of protein folding via fragment folding and assembly. *Protein Sci.* **12** 1177–1187

Hleap JS, Susko E and Blouin C 2013 Defining structural and evolutionary modules in proteins: a community detection approach to explore sub-domain architecture. *BMC Struct. Biol.* **13** 20

Holme P 2005 Core-periphery organization of complex networks. *Phys. Rev. E.* **72** 046111

Hu G, Yan W, Zhou J and Shen B 2014 Residue interaction network analysis of Dronpa and a DNA clamp. *J. Theor. Biol.* **348** 55–64

Hubner IA, Oliveberg M and Shakhnovich EI 2004 Simulation, experiment, and evolution: understanding nucleation in protein S6 folding. *Proc. Natl. Acad. Sci. USA* **101** 8354–8359

Itzhaki LS, Otzen DE and Fersht AR 1995 The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254** 260–288

Jeong H, Mason SP, Barabási AL and Oltvai ZN 2001 Lethality and centrality in protein networks. *Nature* **411** 41–42

Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE and Makse HA 2010 Identification of influential spreaders in complex networks. *Nat. Phys.* **6** 888–893

Kundu S 2005 Amino acid network within protein. *Phys. A.* **346** 104–109

Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T and Ben-Tal N 2005 ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.* **33** W299–W302

Landgraf R, Xenarios I and Eisenberg D 2001 Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **307** 1487–1502

Li H 2009 Predicting protein folding cores based on complex network and phylogenetic analyses. BioMedical Information Engineering, 2009. FBIE 2009. International Conference on Future, IEEE

Li L and Shakhnovich EI 2001 Constructing, verifying, and dissecting the folding transition state of chymotrypsin inhibitor 2 with all-atom simulations. *Proc. Natl. Acad. Sci. USA* **98** 13014–13018

Li J, Wang J and Wang W 2008 Identifying folding nucleus based on residue contact networks of proteins. *Proteins: Struct. Funct. Bioinf.* **71** 1899–1907

Lichtarge O, Bourne HR and Cohen FE 1996 An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257** 342–358

Lin C-C, Juan H-F, Hsiang J-T, Hwang Y-C, Mori H and Huang H-C 2009 Essential Core of Protein-Protein Interaction Network in Escherichia coli. *J. Proteome Res.* **8** 1925–1931

Magyar C, Gromiha MM, Pujadas G, Tusnady GE and In S 2005 SRide: a server for identifying stabilizing residues in proteins. *Nucleic Acids Res.* **33** W303–W305

Newman M 2010 *Networks: An Introduction* (Oxford University Press)

Ng PC and Henikoff S 2003 SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31** 3812–3814

Nicola G and Vakser IA 2007 A simple shape characteristic of proteinâ€'protein recognition. *Bioinformatics* **23** 789–792

Ondrechen MJ, Clifton JG and Ringe D 2001 THEMATICS: a simple computational predictor of enzyme function from structure. *Proc. Natl. Acad. Sci. USA* **98** 12473–12478

Pande VS, Grosberg AY, Tanaka T and Rokhsar DS 1998 Pathways for protein folding: is a new view needed? *Curr. Opin. Struct. Biol.* **8** 68–79

Qin M, Zhang J and Wang W 2006 Effects of disulfide bonds on folding behavior and mechanism of the $\hat{I}^2$-sheet protein tendamistat. *Biophys. J.* **90** 272–286

Riddle DS, Grantcharova VP, Santiago JV, Alm E, Ruczinski I and Baker D 1999 Experiment and theory highlight role of native state topology in SH3 folding. *Nat. Struct. Mol. Biol.* **6** 1016–1024

Seidman SB 1983 Network structure and minimum degree. *Soc. Networks* **5** 269–287

Shander A, Gromiha M, Fawareh H and Sarai A 2004 ASA view: solvent accessibility graphics for proteins. *Bioinformatics* **51** 51

Shen T, Hofmann CP, Oliveberg M and Wolynes PG 2005 Scanning malleable transition state ensembles: comparing theory and experiment for folding protein U1A. *Biochemistry* **44** 6433–6439

Sobolev V, Sorokine A, Prilusky J, Abola EE and Edelman M 1999 Automated analysis of interatomic contacts in proteins. *Bioinformatics* **15** 327–332

Tasdighian S, Di Paola L, De Ruvo M, Paci P, Santoni D, Palumbo P, Mei G, Di Venere A, *et al*. 2013 Modules identification in protein structures: the topological and geometrical solutions. *J. Chem. Inf. Model.* **54** 159–168

Ternstrom T, Mayor U, Akke M and Oliveberg M 1999 From snapshot to movie: analysis of protein folding transition states

taken one step further. *Proc. Natl. Acad. Sci. USA* **96** 14854–14859

Toth-Petroczy A and Tawfik DS 2011 Slow protein evolutionary rates are dictated by surface-core association. *Proc. Natl. Acad. Sci. USA* **108** 11151–11156

Vendruscolo M 2011 Protein regulation: the statistical theory of allostery. *Nat. Chem. Biol.* **7** 411–412

Vendruscolo M, Paci E, Dobson CM and Karplus M 2001 Three key residues form a critical contact network in a protein folding transition state. *Nature* **409** 641–645

Vendruscolo M, Dokholyan NV, Paci E and Karplus M 2002 Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E.* **65** 061910

Vijayabaskar MS and Vishveshwara S 2010 Interaction energy based protein structure networks. *Biophys. J.* **99** 3704–3715

Vishveshwara S, Brinda KV and Kannan N 2002 Protein structure: insights from graph theory. *J. Theor. Comput. Chem.* **1** 187–211

Vishveshwara S, Ghosh A and Hansia P 2009 Intra and intermolecular communications through protein structure network. *Curr. Protein Pept. Sci.* **10** 146–160

Watts DJ and Strogatz SH 1998 Collective dynamics of 'small-world' networks. *Nature*. **393** 440–442

Wuchty S and Almaas E 2005 Peeling the yeast protein network. *Proteomics* **5** 444–449

Yogurtcu ON, Gur M and Erman B 2009 Statistical thermodynamics of residue fluctuations in native proteins. *J. Chem. Phys.* **130** 095103

Zhang J, Zhao H, Xu J-q and Ge X 2009 The *K*-core decomposition and visualization of internet router-level topology. Computer Science and Information Engineering, 2009 WRI World Congress on, IEEE

Zhang H, Zhao H, Cai W, Liu J and Zhou W 2010 Using the *k*-core decomposition to analyze the static structure of large-scale software systems. *J. Supercomput.* **53** 352–369