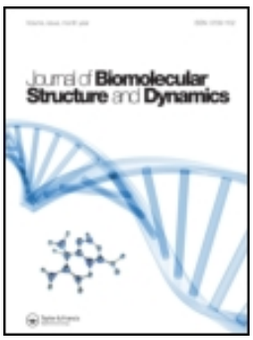

Journal of Biomolecular Structure and Dynamics

ISSN: 0739-1102 (Print) 1538-0254 (Online) Journal homepage: http://www.tandfonline.com/loi/tbsd20

Network properties of protein-decoy structures

Subhojyoti Chatterjee , Moitrayee Bhattacharyya & Saraswathi Vishveshwara

To cite this article: Subhojyoti Chatterjee , Moitrayee Bhattacharyya & Saraswathi Vishveshwara (2012) Network properties of protein-decoy structures, Journal of Biomolecular Structure and Dynamics, 29:6, 1110-1126, DOI: 10.1080/07391102.2011.672625

To link to this article: http://dx.doi.org/10.1080/07391102.2011.672625

Journal of Biomolecular Structure and Dynamics
Vol. 29, No. 6, 2012, 1110–1126

Taylor & Francis
Taylor & Francis Group

Network properties of protein-decoy structures

Subhojyoti Chatterjee, Moitrayee Bhattacharyya and Saraswathi Vishveshwara*

Molecular Biophysics Unit, Indian Institute of Science, Bangalore, 560012, India

Communicated by Ramaswamy H. Sarma

(*Received 1 August 2011; final version received 26 November 2011*)

Convergence of the vast sequence space of proteins into a highly restricted fold/conformational space suggests a simple yet unique underlying mechanism of protein folding that has been the subject of much debate in the last several decades. One of the major challenges related to the understanding of protein folding or *in silico* protein structure prediction is the discrimination of non-native structures/decoys from the native structure. Applications of knowledge-based potentials to attain this goal have been extensively reported in the literature. Also, scoring functions based on accessible surface area and amino acid neighbourhood considerations were used in discriminating the decoys from native structures. In this article, we have explored the potential of protein structure network (PSN) parameters to validate the native proteins against a large number of decoy structures generated by diverse methods. We are guided by two principles: (a) the PSNs capture the local properties from a global perspective and (b) inclusion of non-covalent interactions, at all-atom level, including the side-chain atoms, in the network construction accommodates the sequence dependent features. Several network parameters such as the size of the largest cluster, community size, clustering coefficient are evaluated and scored on the basis of the rank of the native structures and the Z-scores. The network analysis of decoy structures highlights the importance of the global properties contributing to the uniqueness of native structures. The analysis also exhibits that the network parameters can be used as metrics to identify the native structures and filter out non-native structures/decoys in a large number of data-sets; thus also has a potential to be used in the protein 'structure prediction' problem.

Keywords: protein structure networks; non-covalent interaction; network parameters; community size; clique percolation; decoys

1. Introduction

The assembly of proteins into three dimensional structures involves the convergence of a vast sequence space into a highly restricted fold space and the unravelling of the underlying mechanism of protein folding has been one of the grand challenges in biology for several decades. Anfinsen's seminal work on ribonuclease A has established that the protein sequence is programmed to attain a definite three dimensional structure (Anfinsen, 1973), in orders of a few seconds at room temperature (Dill, Ozkan, Shell, & Weikl, 2008). During the past several decades, a large body of experimental and theoretical work has provided insights into the problem of protein folding at different levels. Several reviews/articles have discussed the present status of the protein folding problem (Dobson, 2003; Fersht, 2008; Karplus, 2004). A spirited and timely discussion on the current understanding of protein folding was presented in a series of articles (starting from Sarma, 2011, and the following articles), following the paper on a statistical perspective of protein folding, based on the observed distribution of backbone-level amino acid contacts from a large data-set of protein structures (Mittal, Jayaram, Shenoy, & Bawa, 2010). We briefly describe below some of the approaches and the ideas emerging from the above discussions. In addition, we also present some of the novel ideas such as the network perspective.

At the level of individual amino acids, proline and glycine are known to be involved in punctuating the secondary structures and the hydrophobic residues are involved in the formation of the core (Kauzmann, 1959; Kellis, Nyberg, Sail, & Fersht, 1988) of the protein structure. The role of pair-wise preferential interactions such as disulphide bridges, salt bridges, hydrogen bonding and aromatic stacking in stabilizing various secondary and super-secondary structures has been examined (Burley & Petsko, 1985; Kannan & Vishveshwara, 2000; Kumar & Nussinov, 2001). Furthermore, collective non-bonded interactions among hydrophobic residues leading to the formation of a hydrophobic core (Kellis et al., 1988), the role of charged residues in stabilizing the thermophilic proteins (Berezovsky & Shakhnovich, 2005; Vijayabaskar & Vishveshwara, 2010a) and fold-specific signatures comprised of interacting residues within the protein core have also been identified (Soundararajan,

*Corresponding author. Email: sv@mbu.iisc.ernet.in

ISSN 0739-1102 print/ISSN 1538-0254 online

http://dx.doi.org/10.1080/07391102.2011.672625
http://www.tandfonline.com

Raman, Raguram, Sasisekharan, & Sasisekharan, 2010). While the statistical model (Mittal et al., 2010) has provided a simplistic general picture of the gross features of folded proteins, subtle interplay of preferential interactions has been suggested to account for some of the experimental observations. At a different level, physical principles encompassing the chemical details of the folding phenomenon, such as the thermodynamic factors (Dill et al., 2008; Shakhnovich, 2006), the concept of the folding-funnel and the free energy landscape (Karplus, 2004; Leopold, Montal, & Onuchic, 1992; Onuchic, Luthey-Schulten, & Wolynes, 1997) and the limited geometric space available to the vast amount of protein sequence space, emphasized on the basis of geometry and symmetry (Hoang, Trovato, Seno, Banavar, & Maritan, 2004; Maritan, Micheletti, Trovato, & Banavar, 2000), have also been explored.

The protein structures have also been viewed as networks of connections due to non-bonded interactions between the sequentially separated amino acids (Atilgan, Okan, & Atilgan, 2010; Bagler & Sinha, 2005; Vishveshwara, Ghosh, & Hansia, 2009). This approach goes beyond residue level or pair interaction level and provides a global perspective. It has been an attractive model for a number of investigations ranging from the identification of groups of interacting residues important for folding and function, to the characterization of general network properties of protein structures (Kannan & Vishveshwara, 1999; Sukhwal, Bhattacharyya, & Vishveshwara, 2011). The network analysis of non-covalent interactions of amino acid side-chains in proteins and a comparison with the interaction distributions to those derived from random distributions, in the large percolating unit, has shown the existence of statistical distribution (Brinda, Vishveshwara, & Vishveshwara, 2010), a result that is complementary to the work presented on the basis of backbone Cα analysis (Mittal et al., 2010). However the uniqueness of the protein structure seems to be bestowed by the contacts of higher order, which are captured by the network parameters like cliques/communities or high clustering coefficients (CCoes) (Deb, Vishveshwara, & Vishveshwara, 2009). This concept of percolating communities observed on the basis of geometric overlaps in protein structures, seems to exhibit further subtle features, when the chemistry is brought in the form of interaction energies (Vijayabaskar & Vishveshwara, 2010b).

One of the approaches to investigate protein folding is to generate a large number of decoy structures and compare their properties with those of the native structures. Such decoy structures are also expected to facilitate the design of structures from sequences. It is therefore necessary to device techniques that can efficiently discriminate the decoys or intermediates generated during protein structure prediction from their corresponding native structures. Further, an important component of modelling protein structures is to assess the quality of the generated models, for which various algorithms are available. The quality of the modelled structures is generally assessed by comparing structural and/or energetic features of the decoys with the native structures of proteins. For instance, the structure-based features such as the root mean square deviation (RMSD), the hydrogen bonding patterns and the position of conserved residues in decoys have been shown to distinguish the native from the decoys in certain cases. Scoring functions based on accessible surface area and amino acid neighbourhood considerations were also used for the same (Bahadur & Chakrabarti, 2009). Further, the statistically significant preferential interactions between amino acids have formed the basis of development of knowledge-based pair potentials (Jha, Vishveshwara, & Banavar, 2010; Miyazawa & Jernigan, 1999) and the energy between the interacting pairs of amino acids have been shown to differentiate the native structures from the decoys to some extent (Lu & Sknolnick, 2001). Although, several methodologies have been developed to address this problem, there is no single method to assess the quality in an unambiguous manner, highlighting the need for further development in this area. Graph theoretic approaches are used to capture the global properties of the structures and a few such available studies on the native and the decoy structures are summarized in Table 1. The focus of these investigations however has been mainly at the backbone level (Kucukural, Sezerman, & Ercil, 2008; Taylor & Vaisman, 2006; Vassura, Margara, Fariselli, & Casadio, 2009). In this study, we explore the network features in a large number of decoy sets generated by different methods to enumerate (a) the extent of network features exhibited in the decoy structures and (b) validate the network approach to discriminate the decoys from the native structures, from an all-atom perspective. While the global properties of all atom features are captured by the protein structure networks (PSNs), the optimal backbone packing of secondary structures (measured through the main-chain hydrogen bonds [MHB]) (Banavar, Hoang, Maritan, Seno, & Trovato, 2004; Hoang et al., 2004) capture the compact three-dimensional structure. Thus the network parameters and the MHB in unison are likely to be good discriminators of native from decoys. In this study, a combination of the two approaches (PSN to capture the details of side-chain interactions and the main-chain hydrogen bonding to capture the optimal packing of the polypeptide chain to form secondary structure) is applied on a few selected decoy data-sets in order to evaluate the extent of complementarity between them.

An exhaustive investigation of the network parameters on a large number of decoy data-sets evaluated in this study has provided some interesting insights. Network-based statistically relevant cut-offs, typical to the native three-dimensional structural organizations in proteins, have crystallized from the analysis of a large data-set. Certain guiding principles have also emerged to distinguish the native from their corresponding decoys. The enumeration of such properties and the applicability of this formalism, to model and discriminate the native proteins from the decoy data-sets, are detailed in this study.

Table 1. Network studies on non-covalent interactions in protein-decoy structures.

Investigation details	Taylor and Vaisman (2006)	Kucukural et al. (2008)	Vassura et al. (2009)	Present study
Level of study $C_i^\alpha - C_j^\alpha$ distance cut-off for edge identification	Backbone level 6.5–8.5 Å	Backbone level 6.8 Å	Backbone level 8 Å	Backbone and side chain 1.7–4.5 Å between any pairs of atoms of residue i and j
Network parameters evaluated	Path length, CCoe and node betweenness	Degree, total number of edges and CCoe	Average degree, contact order, normalized complexity and network flow	Total number of edges, LClu, degree distribution, communities size(for $k=3$ clique) and CCoe, PSec
Data-set	Forty-three native proteins and their decoys from Levitt (1992)	PISCES database, 1364 decoy set (Wang et al., 2003)	Rosetta decoy set (Gilis, 2004) and CASP 7 data-set (Samudrala & Levitt, 2000)	Proteins and their decoy from different data-sets (Gilis, 2004; Levitt, 1992; Simons et al., 1999) and CASP 7 data-set (Samudrala & Levitt, 2000)
Residues not considered for non-covalent interaction	$i\pm1$, $i\pm2$, $i\pm3$, $i\pm4$	$i\pm1$	$i\pm4$	$i\pm1$
Structure selection criteria	Resolution < 2.2 Å, R-factor < 0.23, sequence identity < 30%	Resolution < 2.2 Å, R-factor < 0.23, sequence identity < 30%	Decoy structure close (< 0.4 nm) to native structure	Length of the native protein $\geqslant 100$, resolution of the native protein $\leqslant 3$ Å, R-factor < 0.250, equal number of residues for native and decoy proteins

2. Methods

2.1. Selection and cleaning of the decoy data-sets

A wide range of decoy sets are available in the literature. The sets with the native protein of sequence length $\geqslant 100$, obtained from X-ray crystallographic resolution $\leqslant 3$ Å, R-factor lower than 0.25 have been selected. Further, the decoys with the same number of residues as that of the native protein (this criterion provides the basis for uniform normalization of the evaluated network parameters), were selected for our investigation. The decoy sets satisfying these criteria from the following databases were selected.

(a) Data-set from University of Libre de Bruxelles (http://babylone.ulb.ac.be/decoys) (Gilis, 2004).
(b) Data-set from the Baker laboratory (http://depts.washington.edu/bakerpg/) (Simons, Bonneau, Ruczinski, & Baker, 1999).
(c) Data-set from decoys 'R' Us (http://dd.compbio.washington.edu/download.shtml) (Levitt, 1992; Samudrala & Levitt, 2000).
(d) CASP 3 and 7 (predictioncenter.org/casp3/ or predictioncenter.org/casp7/) (Moult et al., 2007; Simons et al., 1999).

The details of the selected proteins and their decoys are presented in Table 2 and Supplementary Table SI.

The Protein Data Bank (PDB) files that were downloaded from various decoy repositories were first cleaned to make them suitable for PSN representation. This involved processes such as handling of missing residues in the native structures (the PDB's were renumbered from 1 leaving out the missing residues) and multiple occupancies, renumbering of the amino acid residues and so on.

2.2. Construction of PSN

A global perspective of the non-covalent side-chain interactions in proteins can be obtained from Protein Structure Network/Protein Structure Graph (PSN/PSG). Details of construction and analysis of PSNs are given elsewhere (Bhattacharyya, Ghosh, Hansia, & Vishveshwara, 2010; Kannan & Vishveshwara, 1999; Sukhwal et al., 2011) and a brief description is given here. The amino acid residues in the structure, with all its side-chain atoms, are considered as 'nodes' and the non-covalent interactions between them are considered as 'edges' for the construction of PSNs from an all-atom perspective. Here, we have quantified these non-covalent side-chain interactions by defining an interaction parameter (I_{ij}) as given below:

$$I_{ij} = (n_{ij}/(\sqrt{N_i \times N_j})) \times 100 \quad (1)$$

where, I_{ij} = strength of interaction between residues i and j (sequence neighbours $i\pm1$ are excluded); n_{ij} = number of distinct interacting atom pairs between i and j (within a distance cut-off of 4.5 Å) and N_i and N_j = normalization values for residues i and j obtained from a statistically significant data-set of proteins, based on the maximum

Table 2. Data-sets selected for the present network study of decoy protein structures.[a]

Decoy database	PDB ID [number of decoy structures]
Set [I]-(Std and Comp): standard and complete collection of decoy data-set (http://babylone.ulb.ac.be/decoys)	1ERV[b], 1FKF, 1LHM, 2RHE, 2RNT. [Number of decoy structures = 200 each]
Set [II]-(Rosetta): Rosetta protein-decoy data-set (http://depts.washington.edu/bakerpg/)	1ACF, 1AIU, 1BKR, 1CG5, 1DHN, 1E6I, 1ELW, 1EW4, 1EYV, 1FKB, 1IIB, 1KPE, 1LOU, 1RNB, 1TUL, 1VLS, 2CHF, 4UBP and 256B. [Number of decoy structures = 120 each]
Decoy R Us data-set (http://dd.compbio.washington.edu/download.shtml)	
Set [III]-(Single): single decoy set	1BP2, 1LH1, 1P2P, 1REI, 1RHD, 1RN3, 2CDV, 2CVP, 2F19, 2I1B, 2ILB, 2PAZ, 2SSI, 2TMN, 2TS1, 3RN3 and 5FD1. [Number of decoy structures = 1 each]
Set [IV]-(HGStruct): HG structural set	1ASH, 1BAB-B, 1COL-A, 1CPC-A, 1EMY, 1FLP, 1GDM, 1HBG, 1HLB, 1HLM, 1HSY, 1ITH-A, 1LHT, 1MBA, 1MYG-A, 1MYJ-A, 1MYT, 2LHB, 2PGH-A, 2PGH-B and 4SDH-A. [Number of decoy structures = 29 each]
Set [V]-(IGStruct): IG structural set	1ACY, 1BAF, 1BBD, 1BBJ, 1DBB, 1DFB, 1DVF, 1EAP, 1FAI, 1FBI, 1FGV, 1FIG, 1FLR, 1FOR, 1FPT, 1FRG, 1FVC, 1FVD, 1GAF, 1GGI, 1GIG, 1HIL, 1HKL, 1IAI, 1IBG, 1IGC, 1IGF, 1IGI, 1IGM, 1IKF, 1IND, 1JEL, 1JHL, 1KEM, 1MAM, 1MCP, 1MFA, 1MLB, 1MRD, 1NBV, 1NCB, 1NGQ, 1NMB, 1NSN, 1OPG, 1PLG, 1RMF, 1TET, 1UCB, 1VFA, 1VGE, 1YUH, 2CGR, 2FB4, 2FBJ, 2GFB, 3HFL, 3HFM, 6FAB, 7FAB and 8FAB. [Number of decoy structures = 60 each]
Set [VI]-(IGHStruct): IG structural hire set[c]	1DVF, 1FGV, 1FLR, 1FVC, 1GAF, 1HIL, 1IND, 1KEM, 1MFA, 1MLB, 1NBV, 1OPG, 1VFA, 1VGE, 2CGR, 2FB4, 2FBJ, 6FAB, 7FAB and 8FAB. [Number of decoy structures = 19 each]
Set [VII]-(CASP 3 and 7): CASP 3 and 7 predicted data-set	1BL0 (973)[d], 2G3 V (458), 2GZV (531), 2H2 W (476), 2HH6 (541), 2HNG (608) and 2IB0 (526)

[a]The details such as the protein name, the number of residues, the resolution, etc. are given in Table S1.
[b]999 decoy structures for the given native protein.
[c]The native proteins are a subset of the IG structure, however the number of decoys are different.
[d]The number in brackets ( ) denotes the number of decoys for the given native proteins.

interaction the residue is capable of making. The interaction strength I_{ij} is evaluated for all pairs of amino acids in the selected protein. A user defined threshold value of I_{ij}, defined as $I_{\min}$, is used to establish an edge between the nodes i and j. The I_{ij} value in proteins generally varies from 0 to 16% (0% represent any single atom–atom contact between the residues i and j and typically no clusters of significant size are obtained above 16%).

The choice of I_{ij} depends on the problem of investigation, for instance, protein–protein interactions can be characterized by strongly interacting (corresponding to $I_{\min}$ of about 6%) interface clusters (Brinda & Vishveshwara, 2005) and the percolation features can be examined at an $I_{\min}$ around 0% (Deb et al., 2009). Also, in the present study, we have characterized the graph features of the native proteins and compared them with those of the decoy structures at $I_{\min}$ around 0%, which provides maximum non-bonded connectivity. Thus, the PSNs have been constructed for all the native proteins and their corresponding decoys listed in Table 2, to analyse their graph features.

2.3. *Graph properties*

The graphs are constructed at $I_{\min}$ of $\geqslant 0\%$ and their features are analysed through a number of parameters ranging from the simple pair-wise interaction (number of edges) to higher order connections like cliques/communities. Specifically, the following parameters are evaluated from PSNs: (a) Number of non-Covalent interactions (NCov), (b) the degree of a node (D), is defined as the number of connections made by a given residue, Degree Distribution (DDis) is the distribution of nodes with different degrees, (c) the largest cluster (LClu) is defined as the set of connected nodes involving the largest number of residues, which represents the bond percolation in PSN and this is evaluated using the Depth First Search algorithm (Cormen, Leiserson, Rivest, & Stein, 2001) and (d) at the higher order connectivity level, cliques/communities are evaluated. A k-clique is defined as a set of k_n nodes, in which all the nodes are connected to each other (Figure 1(a)). In any network (ranging from social

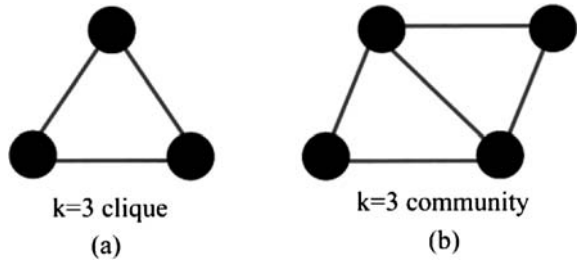

Figure 1. A schematic representation of (a) $k = 3$ clique and (b) a community of four nodes formed by two $k = 3$ cliques sharing $k - 1$ edges.

to protein structures networks), the overlapping and nested architecture is captured by the largest clique/community, which represents the percolation of a higher order connection. The *k*-clique percolation communities can be obtained by rolling a *k*-clique template along the network, thus revealing their highly connected and overlapping features (Palla, Derenyi, Farkas, & Vicsek, 2005). A *k*-clique community is defined as the combination of small *k*-cliques that share node/s. It is the union of a number of *k*-cliques that can be percolated through a series of adjacent *k*-cliques. In the present study, we have used the definition of the *k*-clique community as the one in which two adjacent *k*-cliques share $k-1$ nodes (Figure 1(b)). The cliques/communities are evaluated using the program Cfinder (Adamcsek, Palla, Farkas, Derenyi, & Vicsek, 2006). Generally, $k=3$ cliques are common and communities consisting of *k*-clique with *k* greater than 3 are rare in PSNs, even at the low interaction strength of $I_{\min}=0\%$. The large communities represent the percolation of highly connected units in PSNs. The community with the largest number of nodes is defined as the top community and its size is denoted as 1ComS, similarly the size of second and third largest communities are denoted as 2ComS and 3ComS respectively. The community size drastically reduces after these three largest ones, which are not significant from percolation point of view. We also denote the cumulative size of the three large communities as Cumulative Community Size (CComS). A detailed pictorial depiction to explain the structural basis of these network parameters is given as supplementary Figure SI(a–c). (e) CCoe can be defined as a measure of the degree to which the nodes in a graph tend to cluster together. Here we have evaluated local measure for CCoe, which quantifies the extent of completeness of a graph. The CCoe for each node is calculated by:

$$C_i = 2|\{E_{ij}\}|/k_j(k_j-1) \tag{2}$$

E_{ij} is the actual number of edges between the residue *i* and *j* and for node v_i having k_{j-1} neighbours. $k_j(k_j-1)/2$ are the maximum possible number of edges required to make cliques of k_i number of nodes. Here CCoe is calculated on the LClu obtained.

2.4. Protein secondary structure (PSec)

While the network parameters described above elucidate the details of side-chain interactions, it has been pointed out that the optimal packing of the polypeptide chain (Banavar et al., 2004; Hoang et al., 2004; Rose, Fleming, Banavar, & Maritan, 2006) is the hallmark of folded proteins. This optimal packing is achieved by the formation of secondary structures such as alpha-helices and beta sheets. The side-chain network parameters and the secondary structures provide complementary information required to characterize the features of folded proteins. We have evaluated the secondary structures (PSec) of native and decoy proteins using Rasmol (Sayle & Milner-White, 1995).

2.5. Scoring functions

The general performance of any parameter can be evaluated by Z-score, which is defined as:

$$Z = (N_{\mathrm{P}} - D_{\mathrm{P}})/\mathrm{Std}_{\mathrm{P}} \tag{3}$$

where N_{P} is the value of a given parameter of the native protein, D_{P} and $\mathrm{Std}_{\mathrm{P}}$ are respectively the average and standard deviation for the corresponding decoy sets. Positive values of Z-score indicate that the parameter has performed well in distinguishing the native from the decoys and the confidence is high with larger magnitude. However, the relation between the magnitude and the confidence level is dependent on the number of decoys and their closeness to the chosen native protein, as will be evident from the following section.

3. Results and discussion

In this paper, we have tested our method on 150 proteins and their decoy sets (Table 2), consisting of seven different data-sets (Std and Comp, Rosetta, Single, HGStruct, IGStruct, IGHStruct, CASP 3 and 7). In the following section, we have characterized the network properties of PSNs of the native structures. In the subsequent section, the network features of PSNs of decoys have been discussed and compared with their corresponding native PSNs.

3.1. Network properties of the native protein structures

The study of the network properties of the native proteins provides a reference scale to assess the quality of the decoy structures. It is important to note that these parameters can be used as filters for any protein-decoy set, since we have derived statistically important information from the present data-set of native proteins.

3.1.1. The size of the largest cluster (SLClu)

The size of proteins (*N*) in the data-set varied from 101 to 435. The evaluation of the SLClu at $I_{\min} \geqslant 0\%$ (at least one atom of residue *i* and residue *j* are within 4.5 Å), showed that most of the residues in the native proteins are connected through non-covalent interactions, leading to SLClu being very close to that of the size of the protein, with the range of SLClu being 75–417. This is evident from Figure 2(a). Further, we have obtained the statistics of the deviation of SLClu with the size of the protein in the data-set, as shown in Figure 2(b). It can be seen that in most of the cases, the difference between SLClu and the protein size (*N*) is less than 12 residues (The maximum difference of 28 residues is seen in one protein). Thus we suggest that SLClu smaller than that $(N-11)$ in decoys may imply that these structures

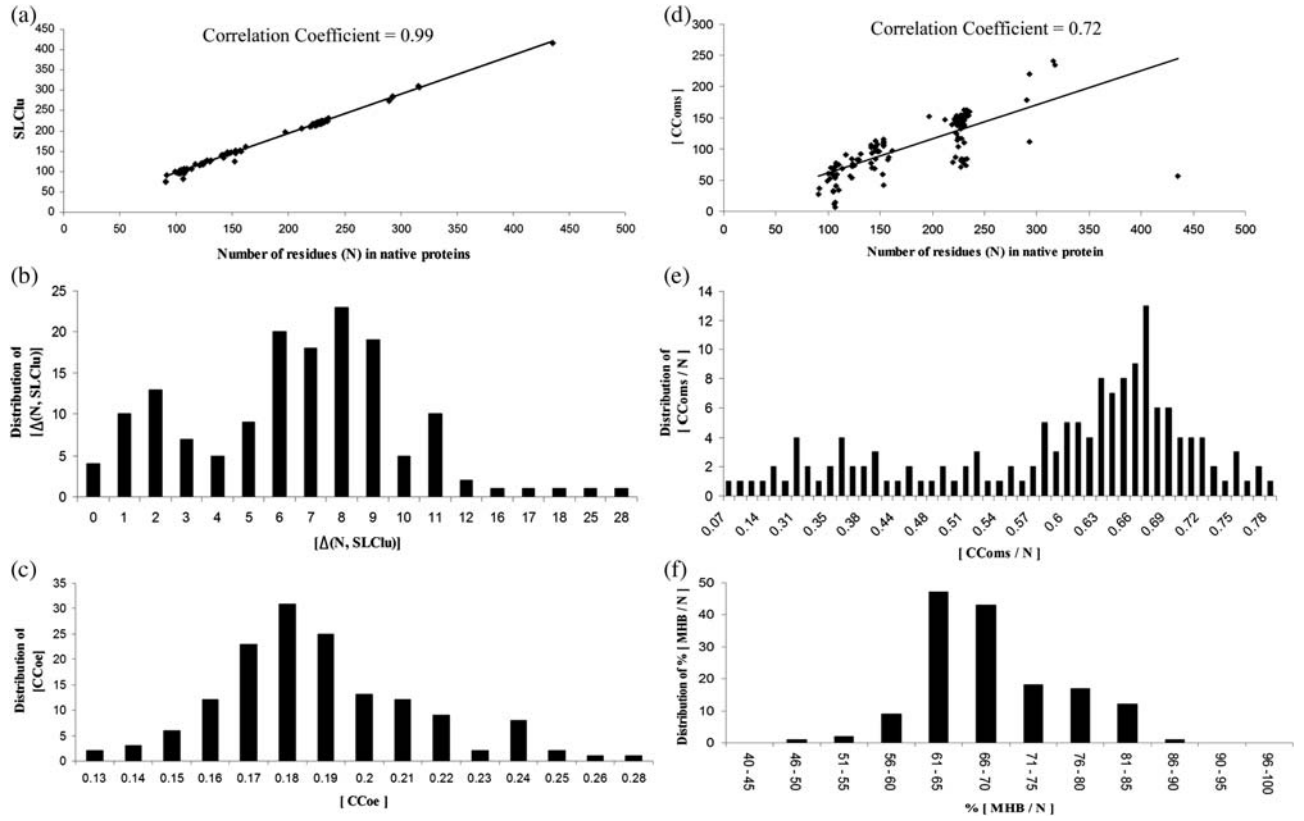

Figure 2. (a) A correlation plot between the number of residues (N) and the SLClu in the native proteins from the data-set given in Table 2, (b) a histogram of the deviation for the size of the largest cluster [Δ(N, SLClu)] from the size of the protein in native structures, (c) a histogram for the distribution of CCoe of the LClu in native structures, (d) a correlation plot between the number of residues (N) and the CComS in the native proteins, (e) a histogram of the fraction of the CComS with respect to the size of the native protein (N) and (f) A histogram of the percentage of the MHB with respect to the size of the native proteins (N).

are away from the native protein. Among the decoys of 150 native proteins, SLClu was between N to $(N-11)$ and $<(N-11)$ in 10,731 and 2127 number of decoys respectively (the details arising from different data-sets are given in Table 3). This amounts to the filtration of only about 17%, with major contribution from CASP 7 (13%) and 2% contribution from the selected data-sets, Std and Comp.

3.1.2. *CCoe of the LClu*

The LClu represents the connectivity at the non-covalent bond level. CCoe provides an idea of the higher level connectivity among the nodes in the LClu. The CCoe values are evaluated for all the native proteins and a histogram of the distribution of this value is depicted in Figure 2(c). The plot shows that majority of the structures have CCoe in the range of 0.15–0.24. A value of CCoe less than 0.15 indicates that the level of connectivity in the decoy is lower than what is normally seen in native proteins. Hence a filter of CCoe < 0.15 is used to identify the decoys as non-native structures. Use of this cut-off value has resulted in detecting 36% of decoys as non-native-like. In this case also, the discriminated decoys come from CASP 7 and the data-set from Std and Comp.

3.1.3. *Cumulative Community Size (CComS)*

Large communities are the percolating units of higher order connectivity. As mentioned in Section 2, the community size is significant only up to three communities and it dwindles drastically after the third community. The parameter CComS captures the total number of residues in top three communities and is correlated with the protein size (Figure 2(d)). A histogram (Figure 2(e)) of CComS/N for this native data-set shows that CComS in majority of native proteins is between 55 and 75% of the size of the protein. Thus a typical protein may be expected to have the community percolating unit of the size greater than 50%. About 42% of decoys may be considered as non-native-like by using a CComS cut-off of 40% (Table 3). Considerable fractions of decoys are filtered from all the data-sets, with the exception of IG structural Hire data-set.

Use of native threshold values for all the three parameters (SLClu, CCoe and CComS) results in the filtration of 32% of decoy structures considered from various data-sets. It should be noted that we have used

Table 3. The number of decoys with network parameters away from the native range (selection criteria from Section 3.1).

Network parameters	Std and Comp (5 native structures with 1799 decoys)	Rosetta (19 native structures with 2280 decoys)	Single (17 native structures with 17 decoys)	HGStruct (21 native structures with 609 decoys)	IGStruct (61 native structures with 3660 decoys)	IGHStruct (20 native structures with 380 decoys)	CASP 3 and 7 (7 native structures with 4113 decoys)
(1) ($\Delta L = [N, SLClu]$)[a]							
No. of decoys with $\Delta L \geqslant 10$	248	7	8	47	82	34	1699
(2) CComs[a]							
No of decoys 2863	with $\leqslant 40\%$ N	1417	241	7	197	673	55
(3) CCoe[a]							
No of 2756	decoys $\leqslant 0.15$	1097	685	7	106	16	7

[a]The terms *N*, SLClu, CComs and CCoe are explained in Section 3.1.

statistically derived threshold values and the methods of generation of the decoys vary extensively. However, individual proteins may deviate from these network parameters. Hence the network properties of the native proteins and their decoys are individually compared in the following section.

3.2. Network properties of decoy structures

Capturing the uniqueness of protein structures has remained a challenge because of the extensive energy landscape available for proteins. Although a large number of decoy sets are available in the literature to test different parameters, they form only a tiny fraction of the possible structures in the conformational space. The parameters developed for the discrimination of the native from the decoy structures should be sensitive and robust to perform well on as many cases as possible, considering the diversity in the native structures and in the methods of decoy structure generation. For instance, the decoys are generated by methods such as homology modelling, simulation, swapping of the main chain and so on (Samudrala & Levitt, 2000).

The present work addresses the details of the network parameters evaluated at all-atom interaction level, including the side chains. (To the best of our knowledge, the side-chain interactions are not included in any of the PSN studies addressing the decoy discrimination problem.) The performance of these parameters at the individual protein level with respect to their decoy sets are explored in the following section.

3.2.1. The number of non-covalent bonds (NCov)

Non-covalent interactions play a dominant role in the integrity of protein structures. Under the constraints of optimally packed backbone secondary structural framework and the excluded volume, the non-covalent interactions are predominantly through side-chain atoms, which accounts for the chemistry of amino acids in the structural context. It is shown in CASP 7 and a few other data-sets that NCov is maximal in the native protein (Bahadur & Chakrabarti, 2009). Here we have evaluated NCov in the data-set of the native and the decoy structures given in Table 2. In the present study, we have enforced the constraints of protein sequence length greater than 100. In smaller proteins (<100), the core to the exposed residue ratio is small and hence their interaction patterns may be different from a typical protein domain. Furthermore, it is also important to consider the native and the corresponding decoys of the same length, as we have done in this study, to avoid errors due to improper normalization of the parameter.

Based on the NCov values, the ranks of the native structures (Figure 3(a)) and the Z-scores (Figure 3(b)) are presented for all the seven data-sets. The rank of the native is high (a low value is considered as high rank and vice versa) in a large number of proteins in data-sets I, III, IV, VI and VII. However, the native ranks are low

for set II (Rosetta) and set V (IG structural). This is also reflected in the Z-score plot, with more negative Z-score cases in sets II and V. From the studies on our data-set, we observe that the native protein occupies the top position with their NCov value being the highest only in 36 (out of 150) proteins. We can observe from Figure 3(a) that the performance of NCov is data-set specific with most of the proteins of Std and Comp, Single and HGStruct data-sets (Table SI), exhibiting high rank. It should also be noted that in our example of IGStruct with 61 proteins (60 decoys each) and IGHStruct with 20 native proteins (19 decoy each), NCov has not discriminated the decoys well. Perhaps this is due to a large number of loops in the IGStruct proteins and spurious interactions in the loops might have contributed to NCov in the decoy structures. The reason for the failure of NCov to distinguish the native in the Rosetta set is not clear.

A large positive Z-score is generally considered to discriminate clearly the native from their decoys. However, the magnitude depends on the number of decoys and difference in the value of the parameter with respect to that of the native reference state. Specifically, the investigation of the Z-score sensitivity as a function of the number of decoys in the data-set shows that the Z-scores for decoys below 100 structures in the data-set are higher and the value stabilizes when the data-set contains around 200–600 structures (details are provided in Figure SV). Additionally, the optimal Z-score values for the parameters NCov and CCoe show a significant dependence on the size of the decoy data-sets as compared to SLClu and CComS, with less significant variations (Figure SV). In our data-set, we find that a Z-score greater than 1.5 discriminates most of the decoys from the native. From Figure 3(b), we see that the Z-score performance is well correlated with the rank profile for NCov, with most of the proteins of Rosetta and some of the cases of IGStruct exhibiting negative Z-score.

While the NCov is the total number of bonds, the degree distribution (DDis) provides the number of nodes with different degree. Generally this distribution varies from 0 to 20 for native proteins as seen from our studies on 150 proteins, with the maximum distribution being around 2–8. However, the values beyond eight are of interest since they are highly connected nodes. A comparison of the degree distribution of the native with the decoys should indicate the pattern preferred by the native structures. Indeed in many cases the nodes with higher degree are greater in the native proteins, compared to their decoys. However, DDis also show data-set dependent features. For instance, the sets which performed well by NCov as mentioned above show that the nodes with degree eight and above are higher for the native as shown for the example cases (Figure 4(a–f)). Rosetta data-set has again failed to discriminate by this measure, with most of the native proteins having smaller number of nodes with high degree in comparison with their decoys (an example case Figure 4(b)). On the other hand, the IGStruct and IGHStruct sets also did not perform well with NCov, but the DDis profiles are mixed (Figure 4(e–f)).

3.2.2. Size of the Largest Cluster (SLClu)

Unlike NCov, which represents pair-wise interaction, SLClu (large number of connected nodes) represents the global property representing the bond-percolating unit of the PSN. The native ranks and the Z-scores are presented respectively in Figure 3(c) and (d). In more than 90% of the cases, the native rank is less than 10 and this is in striking contrast to the NCov where the range for the rank of the Rosetta and IGStruct is as high as 30–60. The fact that the native rank is very high in most cases indicates that the large bond-percolating unit is a property of the native, which is not achieved to the maximum extent in many decoys. Furthermore, this parameter has captured the native of CASP 3 as the top rank and it occupies 16th rank by NCov. Considering the fact that the evaluation is against a large number of decoys (973), this discrimination by both the parameters (SLClu and NCov) is impressive. The Z-score approximately reflects the rank. The Rosetta set has fewer numbers of negative Z-score for SLClu compared to that of NCov. However this improvement is not seen in IGStruct. Here it is important to note that several decoys are also close to or equal to the rank of native, influencing the Z-score (when the RMSD between the native and the decoys is small, often the Z-scores are negative [Tables SIII and SIV]).

3.2.3. CCoe of the LClu

The CCoe represents the extent of connectivity in the LClu in terms of the fraction of edges required to form a clique by a set of nodes belonging to the LClu. This parameter and CComS (discussed in the next section) indeed represent the percolation of higher order connectivity. The value of CCoe in PSNs (native) ranges from 0.1321 to 0.2753 with a mean value of 0.1975. The native rank and the Z-scores are presented in Figure 3(e) and (f). The native rank is below 10 in more than 90% of the cases, as was seen for SLClu. However, there is a great improvement in the performance with drastic reduction in the negative Z-scores. Interestingly, this parameter has distinguished the Rosetta set much better than NCov or SLClu, with negative Z-score for only two cases. The performance of the IGStruct has also improved, however it is not as striking as in the Rosetta set.

3.2.4. Cumulative Size of the top Communities (CComS)

As mentioned in Section 3.1.3, the top three communities contribute to the higher order percolating unit in PSNs. Such communities were shown to contribute significantly to the uniqueness of protein structures and ran-

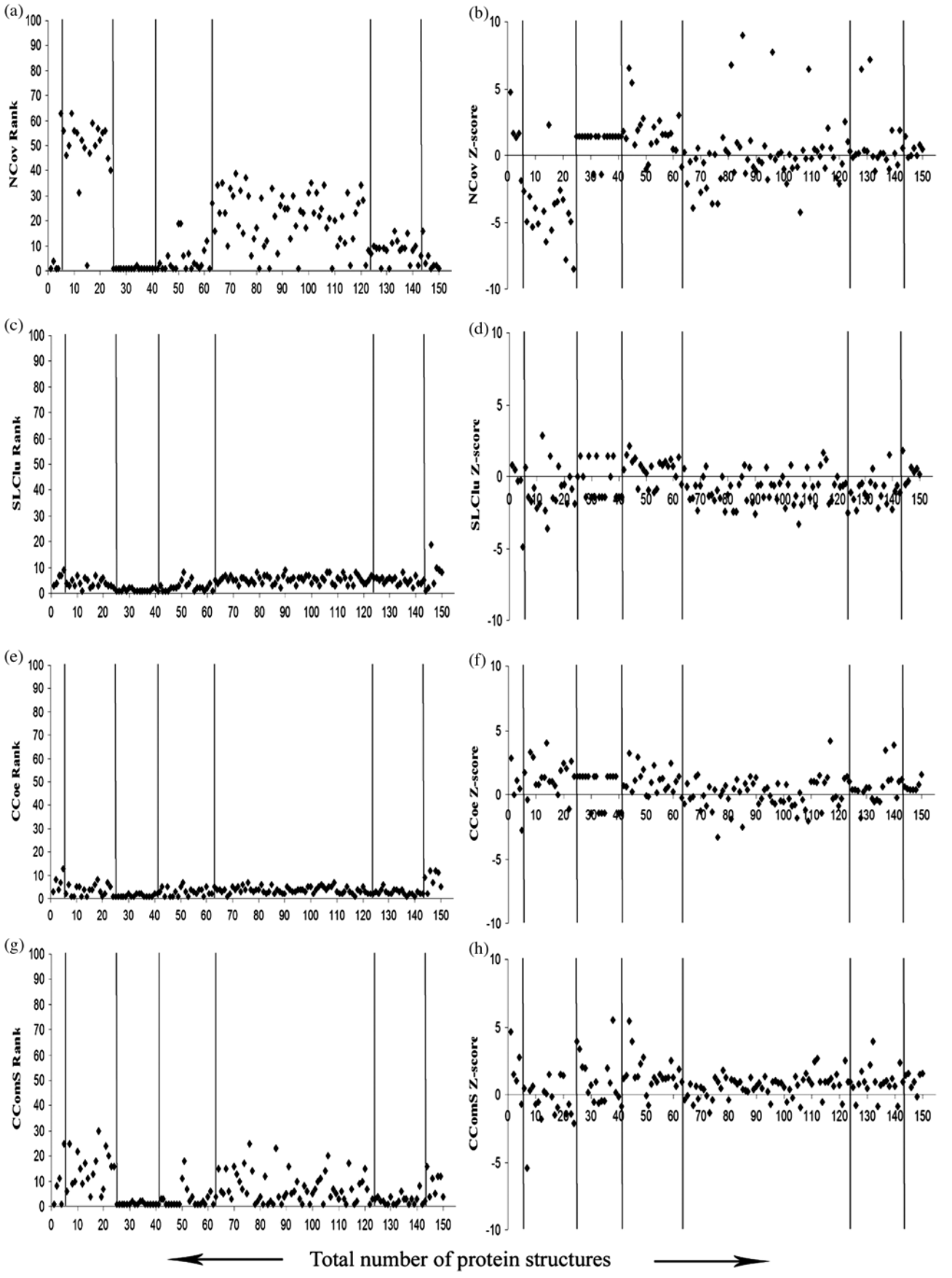

Figure 3. A scatter plot of the network parameters representing rank of the native proteins and the Z-scores for the data-set given in Table 2. The numbers along the *X*-axis correspond to the data-sets Set I [1–5], Set II [6–24], Set III [25–41], Set IV [42–62], Set V [63–123], Set VI [124–143] and Set VII [144–150]. *Y*-axis for figures along the left-hand panel is the rank of the native proteins and Z-score on the right-hand panel.

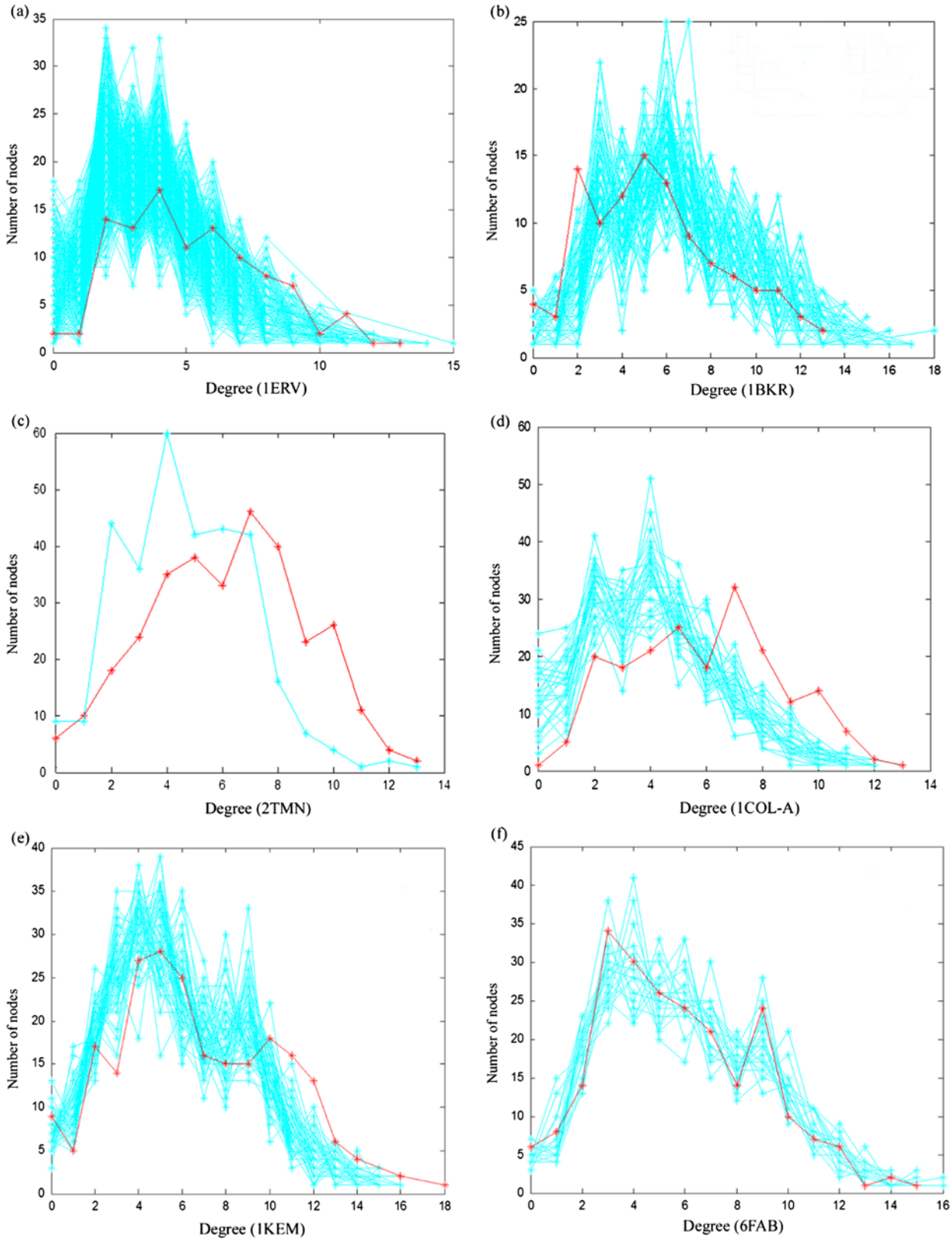

Figure 4. The degree distribution plots for the proteins (a) 1ERV [Std and Comp], (b) 1BKR [Rosetta], (c) 2TMN [Single], (d) 1COL-A [HGStruct], (e) 1KEM [IGStruct] and (f) 6FAB [IGHStruct]. The red curves in these plots correspond to the native and the cyan curves correspond to the decoys.

dom structures were unable to achieve this (Deb et al., 2009). This is the basis behind the evaluation of CComS and comparison between the native and the corresponding decoy structures. While the sizes of the SLClu are

closer to the sizes of the proteins, the community sizes are much smaller. In fact, the top three communities collectively account for about 50% of the protein size. In almost all the cases, the size of the fourth largest community is very small and does not contribute to clique/community percolation. The importance of top three communities is also evident from the fact that the CComS predicts the native as the top rank in many cases, whereas it was not captured as the top rank in the first two top communities. An example case of 1ERV with 999 decoys presented in Figure 5 shows that the native occupies 45th, 3rd and 1st position in the top communities (Anfinsen, 1973), (1 + 2), (1 + 2 + 3) [CComS] respectively. The CComS rank and the Z-scores are shown in Figure 3(g) and (h). The native proteins have ranks ranging from 1 to 30, again the ranks being low for the Rosetta and the IGStruct. However the Z-score has performed well on the Rosetta and the IGStruct sets. Moreover, the negative Z-score for CComS is the least for the IGStruct set, in comparison with other network parameters.

The evaluation of the parameters for CASP (3 and 7) has provided interesting results. We have considered 3140 predicted models for 6 proteins from CASP 7. Interestingly, 265 models failed to make any community, which can be easily discriminated. Furthermore in 2092 models, the CComS is less than 40% of the protein size and this can be considered as away from the native.

3.3. *Distinction of folding intermediates (FDs) from native state: validation of the method*

The elucidation of the pathways of protein folding will greatly benefit from a detailed understanding and characterization of the intermediate states involved in the folding process. The major hurdle in obtaining a high resolution structural insight for a FD lies in their transient appearance during the extremely fast process of kinetic folding. However, with the advancement of NMR techniques, solution state structures of FDs have been solved for a few proteins, namely apocytocrome b_{562}, ribonuclease H, barnase, T4 lysozyme, engrailed homeodomain and Rd-apocytocrome b_{562} (Feng, Takei, Lipsitz, Tjandra, & Bai, 2003). We choose the proteins Rd-apocytocrome b_{562} (Rd-apocyt b_{562}) Feng et al., 2003, T4 lysozyme (Bouvignies et al., 2011), FF domain (Korzhnev, Religa, Banachewicz, Fersht, & Kay, 2010) and the engrailed homeodomain (Religa, Markson, Mayor, Freund, & Fersht, 2005) and their corresponding transiently appearing FDs/mutant forms (as summarized in detail in Table 4) to validate our network-based approach to differentiate between native state and its corresponding decoys. In most of the examples, the RMSD between the native and its FDs are below 3 Å and the conformational differences may be elusive from the conventional methods (Figure SII). It is evident from Figure 6 (also from Figure SIII(a–c)) and Table 4 that mostly all the four network parameters (NCov, SLClu, CCoe and CComS) were able to clearly discriminate the native state from its FDs. It is also striking that for at least two out of the four network parameters, the native structure attained the top rank in contrast to the FDs. Thus our network formalism is highly general and can be applied not only to distinguish native from decoys but also from its preceding FDs.

3.4. *MHB and side-chain networks*

Both the optimal packing of the backbone in the form of secondary structure (Hoang et al., 2004; Rose et al., 2006) and the extensive connection of the side-chain atoms (Deb et al., 2009) as a network, determine the uniqueness of protein structures. In the preceding section we have extensively discussed the network parameters of side-chain interactions. In this section, we have evaluated the MHB to determine the extent of secondary structures

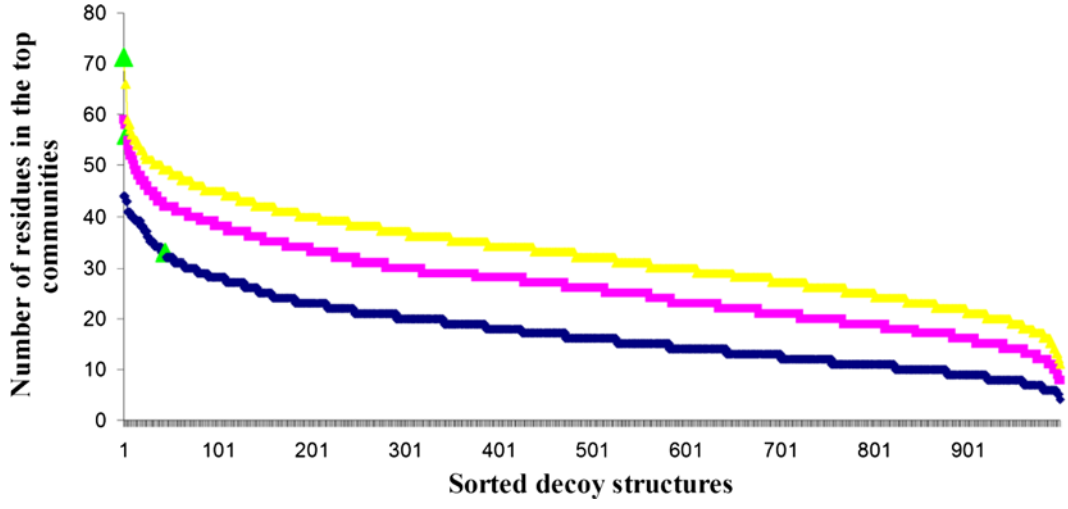

Figure 5. Community size of the decoys for the example [1ERV from Std and Comp set with 999 decoys] data-set. The blue, pink and yellow represent the number of residues in the community [1], [1 + 2] and [1 + 2 + 3] respectively. The positions of the native protein in the corresponding communities are shown by green triangles.

Table 4. Summary and validation of the four network parameters to distinguish between the native and their corresponding FDs for four proteins.

Name of the protein (PDB ID of native – no. of model structures)	No. of FDs	Experimental method for native/intermediates	PDB ID of intermediates (no. of model structures)	Parameters distinguishing well-native from intermediates	RMSD between native and FD (in Å)[a]
Rd-Apocyt b_{562} (1YYJ-10)	2	NMR/NMR	1YZA (Kumar & Nussinov, 2001), 1YZC (Kumar & Nussinov, 2001)	NCov, SLClu, CCoe and CComS	2.58, 3.04
Engrailed Homeodomain (1ENH-1)	1	X-ray/NMR	1ZTR (Sukhwal et al., 2011)	CCoe and CComS	8.51
FF Domain (1UZC-23)	1	NMR/NMR	2KZG (Kumar & Nussinov, 2001)	NCov, SLClu, CCoe and CComS	3.14
T4 Lysozyme (2LZM-1)	2	X-ray, X-ray/NMR	2LC9 (Kumar & Nussinov, 2001), 2LCB (Kumar & Nussinov, 2001)	NCov, CCoe and CComS	0.38, 0.38

Note: Most of the network parameters are able to distinguish well between the native and the corresponding FDs (CCoe and CComS works particularly well).

aThe RMSDs reported in this column are between the first model of the NMR structures/the X-ray structures of the native and the first model of the NMR structures for the FDs respectively.

for selected native proteins and their decoys. About 50–90% of the protein residues adapt secondary structures, as seen from the distribution of MHB in native proteins (Figure 2(f)). Hence the decoys with less than 50% MHB can be detected as adopting non-native structures. However, the behaviour of individual proteins and their decoys may be best captured by Z-scores. We have evaluated the Z-scores for MHB on some selected proteins and their decoy sets and compared their performance with those of network parameters. The results are presented for two sets: (a) a representative from different data-sets, where all Z-scores of the network parameters are positive and at least one of them is greater than +1.5. The results presented in Figure 7(a) show that only in four cases (out of 11), the Z-score for MHB is greater than or equal to 1.5. This emphasizes the fact that the network parameters are clearly able to distinguish the cases which were not captured by MHB, (b) MHB was investigated (Figure 7(b)) for a set of 13 proteins and their decoys in which the Z-score of at least one of the network parameter is positive. Strikingly, the MHB Z-scores are negative for all the cases, indicating that the MHB alone is not able to capture the subtle interactions at the side-chain level that is captured by at least one of the network parameters. The difference in MHB and the network parameters for the single decoy sets is evaluated. Clearly, the CComS has performed extremely well in discriminating the misfolded or the PDB error structures from the native as shown in Table 5. Two interesting examples (2F19 and 2TMN) are plotted in Figure 8(i (a) and (b)) and (ii(a) and (b)). In the case of 2F19, the decoy is a PDB error structure, which was obtained before substantial refinement, thus the RMSD difference is marginal (0.98). This decoy structure is distinguished by a drastic reduction of both MHB and CComS. In the case of 2TMN, the RMSD is extremely high due to misfolding, in which the main chain is swapped in the decoy structure. Although there is no significant difference in MHB, a dramatic difference is seen in CComS, providing support to the concept (Deb et al., 2009) that large communities involving side-chain interactions contribute to the uniqueness of protein structures.

To summarize, to the best of our knowledge, this is the first attempt to consider the side-chain interactions in the context of the identification of the native structures from a set of decoys. We have characterized a large number of native proteins in terms of non-covalent interactions of the side chains. We have investigated several network parameters and the range of their values in native proteins. This can act as a filter to eliminate non-native structures in the context of protein structure selection problem. The parameters have been tested on a large number of decoy sets and we have shown that the performance of network parameters is encouraging, although in some cases it is found to be data-set specific. The network parameters capture the global connectivity, with NCov in terms of the pairwise connections, SLClu in terms of the largest bond

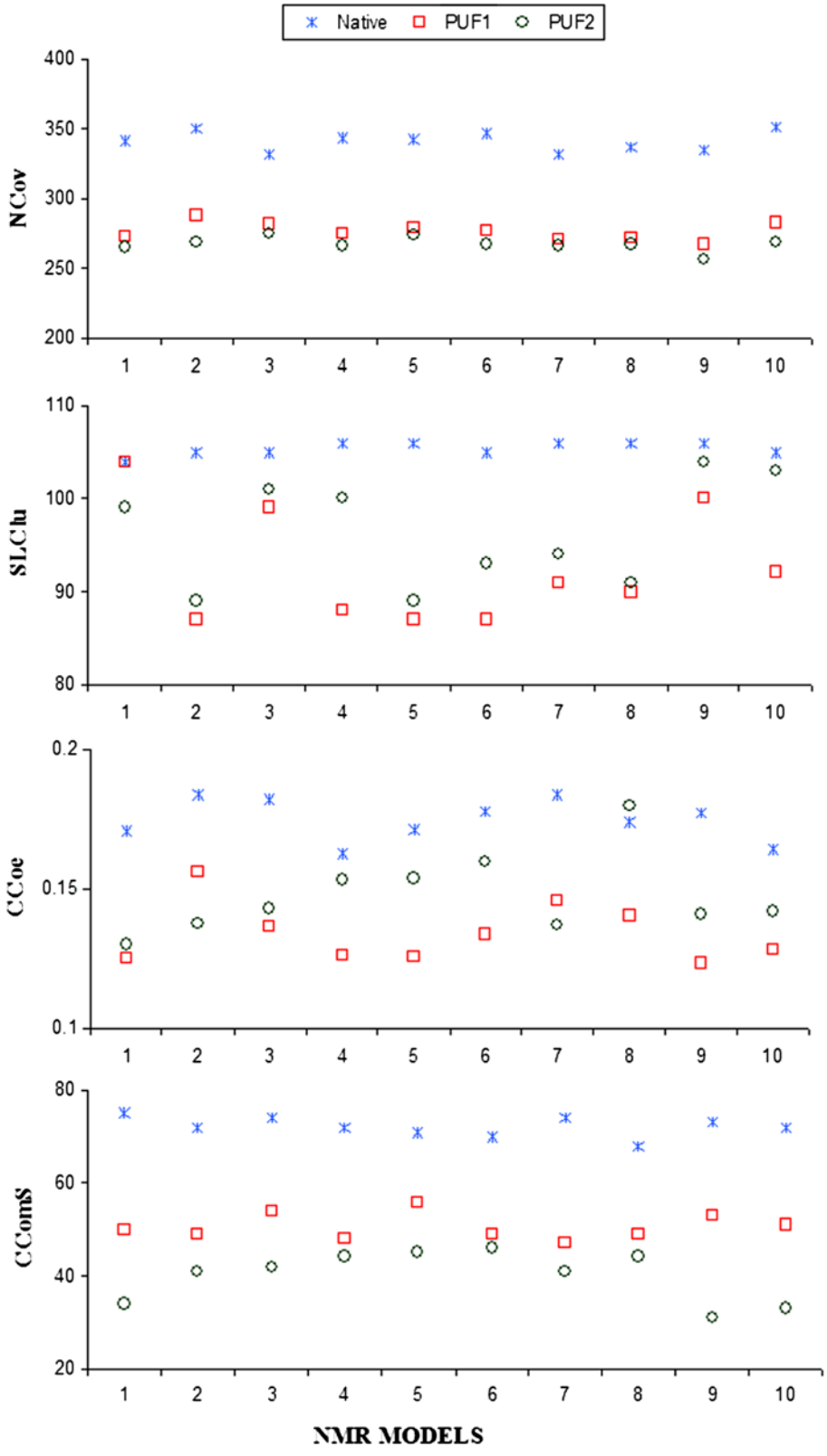

Figure 6. Plot of each of the four network parameters (NCov, SLClu, CCoe and CComS) for the 10 NMR models of the native, and two partially unfolded forms, PUF1 and PUF2 for the protein Rd-apocyt b_{562}. All of the network parameters could distinguish well between the native and the PUFs, with the native attaining the top rank in most of the cases. This shows that the native state is bestowed with unique topological features.

percolating entity, CComS in terms of the largest clique/community (higher order) percolation in proteins. The CCoe of the LClu also measures the extent of connectivity in the LClu. In general, we have used a posi-

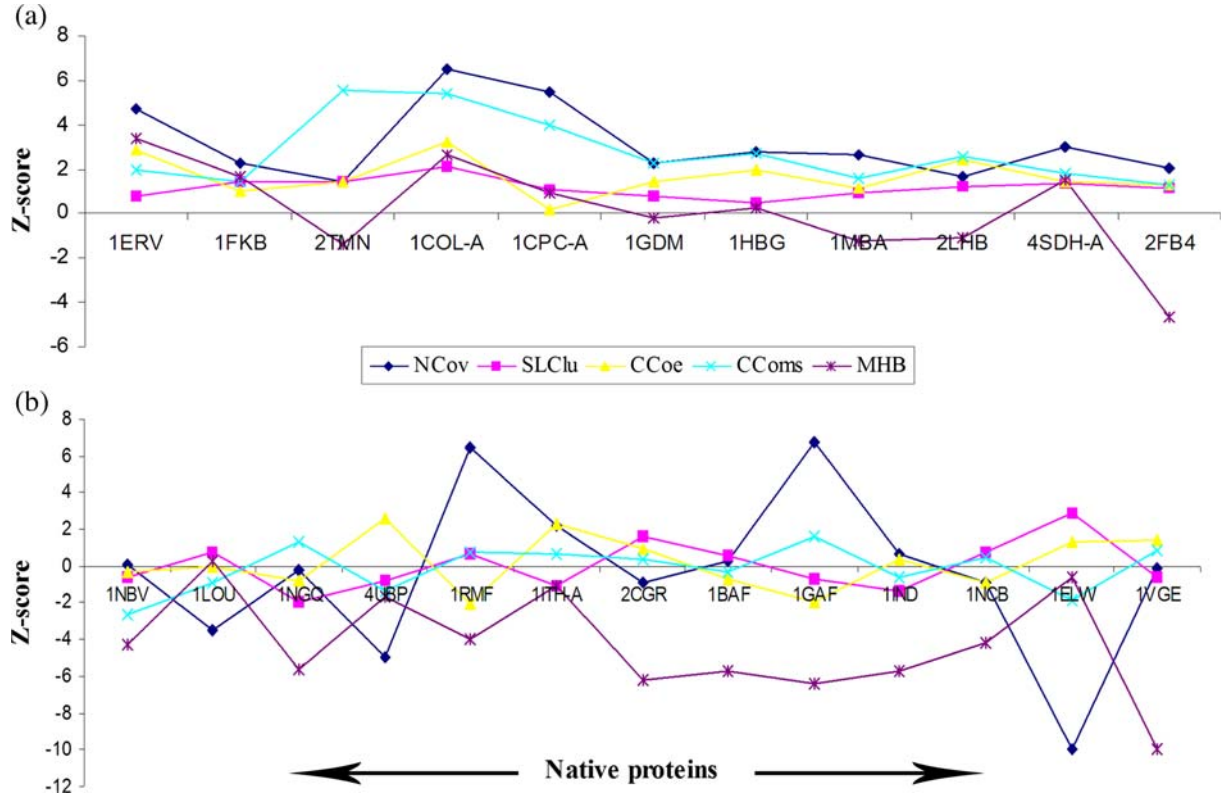

Figure 7. Comparison of Z-scores of the network parameters with MHB representing the secondary structures. (a) The proteins with all network parameters positive and at least one of the network parameter Z-score $\geqslant 1.5$, (b) the proteins with any one network parameter being positive. The protein names are given along the *X*-axis and their Z-scores on the *Y*-axis.

Table 5. Difference of the main-chain hydrogen bond (ΔMHB) and the cumulative community size (ΔCComS) between the native and the decoy proteins [single decoy set].

Serial number	Protein name (PDB ID)	*N*	ΔRMSD	ΔMHB	ΔCComS
1.	Bovine Pancreatic Phospholipase (1BP2)	123	15.38	−5	50
2.	Leghemoglobin (1LH1)	153	16.21	8	56
3.	Porcine Pancreatic Phospholipase (1P2P)	124	18.31	−3	33
4.	Bence-Jones protein (1REI)	212	18.09	7	40
5.	Bovine Liver Rhodanese (1RHD)	293	20.28	−26	27
6.	Bovine Ribonuclease A (1RN3)	124	18.36	−8	11
7.	Cytochrome (2CDV)	107	14.14	−10	−14
8.	Yeast Cytochrome C Peroxidase (2CYP)	293	20.44	3	81
9.	3-D structure of 2-crystal forms of FAB R19.9 (2F19)	435	0.981	112	152
10.	Interleukin-1 Beta (2I1B)	153	16.25	−24	−76
11.	Tetracenomycin Polyketide Synthesis Protein (2ILB)	153	15.25	−24	−24
12.	Pseudoazurin (2PAZ)	123	15.38	1	21
13.	Proteinase Inhibitor Streptomyces Subtilisin Inhibitor (2SSI)	107	14.18	4	12
14.	Phosphoramidates (2TMN)	316	22.44	−1	153
15.	Tyrosyl-T/RNA Synthetase (2TS1)	317	22.5	−14	29
16.	Ribonuclease A (3RN3)	124	18.36	−8	29
17.	Azotobacter Vinelandii Ferredoxin (5FD1)	106	13.21	34	57

tive value of Z-score as a measure of good performance. Only in 16 (out of 150) cases, none of the network parameters have positive Z-score. The CComS and CCoe have done very well with 107 and 100 proteins with positive Z-score respectively (45 cases and 27 cases of the native have the highest rank of CComS and CCoe respectively). The success cases are 83 and 54 respectively for NCov and SLClu. Thus, it is clear that higher order percolation is one of the unique features of native proteins. Performance of multiple parameters is tabulated in the supplementary Table SII. All the four parameters are positive in 34 cases and NCov, CComS and CCoe are positive in 27 cases.

Here we present a critical assessment of the performance of network parameters. As discussed earlier, the Z-score of SLClu has not performed well due to a

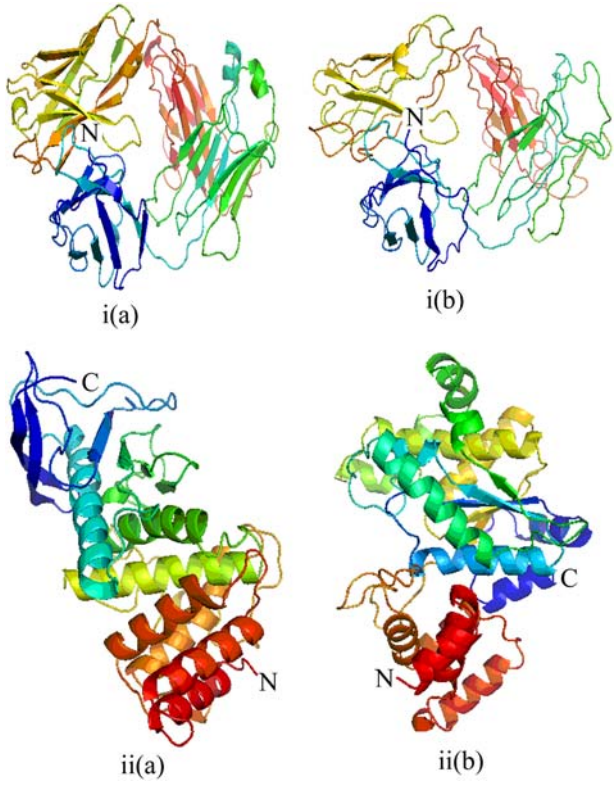

Figure 8. Comparison of native and decoy structures from single decoy sets. (i) 2F19 (a) native, (b) decoy [PDB error structure] in terms of parameters [ΔRMSD = 0.981 Å; ΔSLClu = −3; ΔCComs = 152; ΔCCoe = −0.0325; ΔMHB = 112] and (ii) 2TMN (a) native, (b) decoy [misfolded structure] in terms of parameters [ΔRMSD = 22.44 Å; ΔSLClu = 5; ΔCComs = 153; ΔCCoe = 0.012; ΔMHB = −1]. These plots are generated using the program, Pymol (26), following the *chain* colour scheme.

large number of decoys having near-native values. We can rationalize this observation as the SLClu is usually comprised of almost 90% of the total number of residues (N), both in the native and in the near-native structures, leading to less distinguishing capacity. Other network parameters with very similar values for the native and near-native structures also show low positive Z-scores. A detailed summary of the performance of the four network parameters for near-native and far-native decoy structures is presented in supplementary Table SIII and Figure SIV, for one representative member of each data-set. It is evident that most of the parameters distinguish both the near and far-native structures from the native with reasonably high Z-score values, with few data-set specific exceptions. However, it is worth mentioning that the higher order connectivity parameters like CCoe and CComS often serve as better metrics than NCov or SLClu in near-native cases. Furthermore, when the RMSD between the native and the near-native decoys is small (<2 Å), often the Z-scores are negative (Tables SIII and SIV). In our current study, Rosetta and IGHStruct data-sets, in which most of the decoys are close structural neighbours of the native (low RMSD, usually < 2 Å) (Table SIV), have shown negative Z-scores for several network parameters (Table SIII). This implies that individual network parameters might not be able to distinguish well between native and decoy structures with low RMSD's (i.e. for certain near-native cases). The four network parameters used in this study may be optimally combined by using machine learning techniques (such as support vector machines) to derive a general metric that is universally suited for distinguishing native from decoys. Thus, the detailed network-based methodology designed in this article has a far reaching potential to be applied in protein folding studies and in protein structure prediction for filtering out the native state from the non-native ones. Further, it may be used in Monte Carlo simulations to restrict the sampled conformational space and carry out native-oriented folding dynamics.

The number of MHB is a good measure to evaluate the optimal packing of secondary structures in proteins. On the other hand, the sequence of amino acids imparts uniqueness to the protein structure. This uniqueness is captured to some extent by the large percolating communities formed by the interaction among the side-chain atoms. Thus the higher order network measures such as large communities and CCoe are likely to identify the native structures more effectively than any of the known parameters.

4. Conclusion

The features of networks constructed for about 150 proteins and their decoy structures on the basis of non-covalent interactions have been analysed in this study. The investigated parameters are (a) the number of non-covalent interactions (NCov), (b) the SLClu, (c) the clustering coefficient of the largest cluster (CCoe) and (d) the size of the top large communities (CComS). The investigations on the native proteins provided a statistical range of the network parameters which can be used as measures to identify native structures and discard decoy structures in a systematic manner. However these measures cannot be used in an unequivocal manner, since the diversity of protein structures is vast and the decoys cover only a small fraction of the structure space, although they are generated by multiple methods.

The rank of the native structure in comparison with the decoys and the Z-scores evaluated for the four network parameters have shown good performance, however, is data-set specific. For instance, some of the parameters were not able to identify the native structure from the Rosetta and the IG Structural data-sets. Interestingly, best performance in most of the cases is seen by the network parameters CComS and CCoe, which represent higher order connectivity in the percolating unit along the structure. Further these parameters have also performed well on six CASP 7 data-sets, thus exhibiting

the potential of these measures in evaluating the predicted structures.

The number of MHB, representing the optimal packing of the backbone secondary structures was also evaluated for some selected cases. The discriminatory power of this measure is also good. However it failed in a few cases, where the network parameters were successful, emphasizing the importance of side-chain connectivity along with the optimal backbone packing.

Acknowledgements

We acknowledge the Department of Science and Technology (DST Mathematical Biology Grant), India, for support. We thank Anupam Nath Jha for the discussions on decoys. The concept of higher order percolations in proteins used in this manuscript evolved through discussion with Smitha Vishveshwara of UIUC. Microsoft Research supported the fellowship of SC. MB and SV acknowledge the Council of Scientific and Industrial Research (CSIR), India, for senior research fellowship and emeritus professorship respectively.

Supplementary material

The supplementary material for this paper is available online at http://dx.doi.org/10.1080/07391102.2011.672625.

References

Adamcsek, B., Palla, G., Farkas, I.J., Derenyi, I., & Vicsek, T. (2006). CFinder: Locating cliques and overlapping modules in biological networks. *Bioinformatics, 22*, 1021–1023.

Anfinsen, C.B. (1973). Molecular thermodynamic model to predict the alpha-helical secondary structure of polypeptide chains in solution. *Science, 181*, 223–230.

Atilgan, C., Okan, O.B., & Atilgan, A.R. (2010). How orientational order governs collectivity of folded proteins. *Proteins: Structure, Function, Bioinformatics, 78*, 3363–3375.

Bagler, G., & Sinha, S. (2005). Network properties of protein structures. *Physica A: Statistical Mechanics and its Applications, 346*, 27–33.

Bahadur, R., & Chakrabarti, P. (2009). Discriminating the native structure from decoys using scoring functions based on the residue packing in globular proteins. *BMC Structural Biology, 9*, 76–85.

Banavar, J.R., Hoang, T.X., Maritan, A., Seno, F., & Trovato, A. (2004). Unified perspective on proteins: a physics approach. *Physical Review E, 70*, 041905.

Berezovsky, I.N., & Shakhnovich, E.I. (2005). Physics and evolution of thermophilic adaptation. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 12742–12747.

Bhattacharyya, M., Ghosh, A., Hansia, P., & Vishveshwara, S. (2010). Allostery and conformational free energy changes in human tryptophanyl-tRNA synthetase from essential dynamics and structure networks. *Proteins, 78*, 506–517.

Bouvignies, G., Vallurupalli, P., Hansen, D.F., Correia, B.E., Lange, O., Bah, A., … Kay, L.E. (2011). Solution structure of a minor and transiently formed state of a T4 lysozyme mutant. *Nature, 477*, 111–114.

Brinda, K.V., & Vishveshwara, S. (2005). A network representation of protein structures: Implications for protein stability. *Biophysical Journal, 89*, 4159–4170.

Brinda, K.V., Vishveshwara, S., & Vishveshwara, S. (2010). Random network behaviour of protein structures. *Molecular BioSystems, 6*, 391–398.

Burley, S.K., & Petsko, G.A. (1985). Aromatic-aromatic interaction: A mechanism of protein structure stabilization. *Science, 229*, 23–28.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., & Stein, C., (2001). Introduction to Algorithms, Second Edition. Cambridge, MA: MIT Press, 540–549.

Deb, D., Vishveshwara, S., & Vishveshwara, S. (2009). Understanding protein structure from a percolation perspective. *Biophysical Journal, 97*, 1787–1794.

Dill, K.A., Ozkan, S.B., Shell, S.M., & Weikl, T.R. (2008). The protein folding problem. *Annual Review of Biophysics, 37*, 289–316.

Dobson, C.M. (2003). Protein folding and misfolding. *Nature, 426*, 884–890.

Feng, H., Takei, J., Lipsitz, R., Tjandra, N., & Bai, Y. (2003). Specific non-native hydrophobic interactions in a hidden folding intermediate: Implications for protein folding. *Biochemistry, 42*, 12461–12465.

Fersht, A.R. (2008). From the first protein structures to our current knowledge of protein folding: Delights and scepticisms. *Nature Reviews Molecular Cell Biology, 9*, 650–654.

Gilis, D. (2004). Protein decoy sets for evaluating energy functions. *Journal of Biomolecular Structure & Dynamics, 21*, 725–854.

Hoang, T.X., Trovato, A., Seno, F., Banavar, J.R., & Maritan, A. (2004). Geometry and symmetry presculpt the free-energy landscape of proteins. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 7960–7964.

Jha, A.N., Vishveshwara, S., & Banavar, J.R. (2010). Amino acid interaction preferences in proteins. *Protein Science, 19*, 603–616.

Kannan, N., & Vishveshwara, S. (1999). Identification of side-chain clusters in protein structures by a graph spectral method. *Journal of Molecular Biology, 292*, 441–464.

Kannan, N., & Vishveshwara, S. (2000). Aromatic clusters: A determinant of thermal stability of thermophilic proteins. *Protein Engineering, 13*, 753–761.

Karplus, M. (2004). Behind the folding funnel diagram. *Nature Chemical Biology, 7*, 401–404.

Kauzmann, W.J. (1959). Some factors in the interpretation of protein denaturation. *Advances in Protein Chemistry, 14*, 1–63.

Kellis, J.T., Nyberg, K., Sail, D., & Fersht, A.R. (1988). Contribution of hydrophobic interactions to protein stability. *Nature, 333*, 784–786.

Korzhnev, D.M., Religa, T.L., Banachewicz, W., Fersht, A.R., & Kay, L.E. (2010). A transient and low-populated protein-folding intermediate at atomic resolution. *Science, 329*, 1312–1316.

Kucukural, A., Sezerman, O., & Ercil, A. (2008). Discrimination of native folds using network properties of protein structures. *Advances in Bioinformatics and Computational Biology, 6*, 59–67.

Kumar, S., & Nussinov, R. (2001). How do thermophilic proteins deal with heat? *Cellular and Molecular Life Sciences, 58*, 1216–1233.

Leopold, P.E., Montal, M., & Onuchic, J.N. (1992). Protein folding funnels: A kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences of the United States of America, 89*, 8721–8725.

Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. *Journal of Molecular Biology, 226*, 507–533.

Lu, H., & Sknolnick, J (2001). A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins: Structure, Function, Bioinformatics, 44*, 223–232.

Maritan, A., Micheletti, C., Trovato, A., & Banavar, J.R. (2000). Optimal shapes of compact strings. *Nature, 406*, 287–290.

Mittal, A., Jayaram, B., Shenoy, S., & Bawa, T.S. (2010). A stoichiometry driven universal spatial organization of backbones of folded proteins: Are there Chargaff's rules for protein folding? *Journal of Biomolecular Structure & Dynamics, 28*, 133–142.

Miyazawa, S., & Jernigan, R.L. (1999). Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins: Structure, Function, Bioinformatics, 34*, 49–68.

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Hubbard, T., & Tramontano, A. (2007). Critical assessment of methods of protein structure prediction—Round VII. *Proteins: Structure, Function, Bioinformatics, 69*, 3–9.

Onuchic, J.N., Luthey-Schulten, Z., & Wolynes, P.G. (1997). Theory of protein folding: the energy landscape perspective. *Annual Review of Physical Chemistry, 48*, 545–600.

Palla, G., Derenyi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature, 435*, 814–820.

Religa, T.L., Markson, J.S., Mayor, U., Freund, S.M.V., & Fersht, A.R. (2005). Solution structure of a protein denatured state and folding intermediate. *Nature, 437*, 1053–1056.

Rose, G.D., Fleming, P.J., Banavar, J.R., & Maritan, A. (2006). A backbone-based theory of protein folding. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 16623–16633.

Samudrala, R., & Levitt, M. (2000). Decoys 'R' Us: A database of incorrect conformations to improve protein structure prediction. *Protein Science, 9*, 1399–1401.

Sarma, R.H. (2011). A Conversation on Protein Folding. *Journal of Biomolecular Structure and Dynamics, 28*, 587–588.

Sayle, R., & Milner-White, E.J. (1995). RASMOL: Biomolecular graphics for all. *Trends in Biochemical Sciences, 20*, 374–376.

Shakhnovich, E. (2006). Protein folding thermodynamics and dynamics: Where physics, chemistry, and biology meet. *Chemical Reviews, 106*, 1559–1588.

Simons, K.T., Bonneau, R., Ruczinski, I., & Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins: Structure, Function, Bioinformatics, 37*, 171–176.

Soundararajan, V., Raman, R., Raguram, S., Sasisekharan, V., & Sasisekharan, R. (2010). Atomic interaction networks in the core of protein domains and their native folds. *PLoS ONE, 5*, e9391.

Sukhwal, A., Bhattacharyya, M., & Vishveshwara, S. (2011). Network approach for capturing ligand-induced subtle global changes in protein structures. *Acta Crystallographica Section D: Biological Crystallography, 67*, 429–439.

Taylor, T.J., & Vaisman, I.I. (2006). Graph theoretic properties of networks formed by the Delaunay tessellation of protein structures. *Physical Review E, 73*, 041925.

Vassura, M., Margara, L., Fariselli, P., & Casadio, R. (2009). A graph theoretic approach to protein structure selection. *Artificial Intelligence in Medicine, 45*, 229–237.

Vijayabaskar, M.S., & Vishveshwara, S. (2010). Comparative analysis of thermophilic and mesophilic proteins using Protein Energy Networks. *BMC Bioinformatics, 11*, S49.

Vijayabaskar, M.S., & Vishveshwara, S. (2010). Interaction energy based protein structure networks. *Biophysical Journal, 99*, 3704–3715.

Vishveshwara, S., Ghosh, A., & Hansia, P. (2009). Intra and inter-molecular communications through protein structure network. *Current Protein & Peptide Science, 10*, 146–160.

Wang, G., & Dunbrack, R.L. (2003). PISCES: A protein sequence culling server. *Bioinformatics, 19*, 1589–1591.