Send Orders for Reprints to reprints@benthamscience.ae

Protein Structure and Function: Looking through the Network of Side-Chain Interactions

Moitrayee Bhattacharyya[1,#,$], Soma Ghosh[2,$] and Saraswathi Vishveshwara[1,*]

[1]*Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India;* [2]*I.I.Sc. Mathematics Initiative, Indian Institute of Science, Bangalore, India*

S. Vishveshwara

Abstract: Network theory has become an excellent method of choice through which biological data are smoothly integrated to gain insights into complex biological problems. Understanding protein structure, folding, and function has been an important problem, which is being extensively investigated by the network approach. Since the sequence uniquely determines the structure, this review focuses on the networks of non-covalently connected amino acid side chains in proteins. Questions in structural biology are addressed within the framework of such a formalism. While general applications are mentioned in this review, challenging problems which have demanded the attention of scientific community for a long time, such as allostery and protein folding, are considered in greater detail. Our aim has been to explore these important problems through the eyes of networks. Various methods of constructing protein structure networks (PSN) are consolidated. They include the methods based on geometry, edges weighted by different schemes, and also bipartite network of protein-nucleic acid complexes. A number of network metrics that elegantly capture the general features as well as specific features related to phenomena, such as allostery and protein model validation, are described. Additionally, an integration of network theory with ensembles of equilibrium structures of a single protein or that of a large number of structures from the data bank has been presented to perceive complex phenomena from network perspective. Finally, we discuss briefly the capabilities, limitations, and the scope for further explorations of protein structure networks.

Keywords: Allostery, graph theory, molecular dynamics, protein structure network, sidechain interactions, structure quality assessment, structure prediction, tRNA synthetases.

"The only real voyage of discovery consists not in seeking new landscapes, but in having new eyes" - ***Marcel Proust***

1. INTRODUCTION

1.1. Network Theory and Biology

Systems in biology are complex. This complexity can be at the level of interactions between molecules, genes, cells, tissues, organs, organisms, or communities based on the resolution of the interrogation. The key challenge is to model the behavior of these complex systems. The goal is to find approaches that hit a middle ground between reductionist and holistic approaches to ensure that one can study and probe these systems, but not get lost in the details.

Network theory provides an excellent description of a complex 'interactome'. It involves the study of graphs, a mathematical formulation to represent the connectivity between the constituent elements of a set as edges and nodes. Networks allow us to efficiently model the interactions between the components in a biological system, thus enabling an arsenal of mathematical algorithms for its analyses. Network analysis leads to the characterization of complex biological systems into well-understood theoretical models. Additionally it provides a tunable 'looking glass', where the extent of network complexity can be scaled based on the definition of the nodes and edges constituting the graph/network. These key aspects of network theory have made it the method of choice among the scientific community probing a wide array of relevant biological questions. It ranges from modelling of protein/macromolecular structures, protein-protein interactions, genetic interactions/perturbations-disease relations, signal transduction in the cellular context, application to drug discovery, to societal interaction and spread of disease to name just a few. In the recent years, comprehensive reviews have addressed many of these applications in great details [1]. The application of graph theoretical concepts in the investigation of detailed atomic interactions in protein structures from a global perspective constitutes the main focus of this review. Although earlier reviews [2] had dealt with this subject, sufficient advancement has taken place in the last few years on topics such as allostery and structure validation, which involves the exploration of newer network approaches and investigation of ensembles of structures. The present review focuses on some of these developments and the necessary background is provided below.

*Address correspondence to this author at the Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India; Tel: +91-80-2293-2611; Fax: +91-80-2360-0535; E-mail: sv@mbu.iisc.ernet.in

[#]*Current address: Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, United States*

[$]*Equal contribution*

1875-5550/16 $58.00+.00

1.2. Architecture of Proteins: A Network View

With an explosion in the availability of genomic and structure data, there is an increasing need to understand the relationship between sequence-structure-function of proteins and how information percolates through them. It has been suggested that both the three dimensional (3D) organization of proteins as well as the conformational dynamics encode function. The complex three-dimensional organization of macromolecules emanates from the unique atomic level interactions among its building blocks (amino acids for proteins and nucleotides for RNA/DNA). Such interactions need to be represented and probed from a global perspective in order to understand their influence in the complex packing of proteins. Thus it is imperative to look at these local interactions in the global milieu of the structure.

1.2.1. Nodes: The Building Blocks

In the context of macromolecular structure, the nodes translate to the corresponding building blocks and an edge is drawn between two nodes if they interact with each other. The elements that comprise the nodes in macromolecules range from backbone atoms (e.g., Cα/Cβ atoms in amino acids) [3] to side-chain atoms [2c, 4] or a vector representation of the secondary structures [5]. Further the side chain based networks may be constructed by considering all atoms or may be coarse grained to the geometric center of the side chain atoms [6].

1.2.2. Edges: The Ties

Edges are defined as connections between the nodes in a network. Nodes in a protein structure network are connected in various ways based on user-defined criteria for interactions. A geometry or topology based structure network of non-covalent interactions in proteins may be obtained in cases where edges in the simplest form are described by a cut-off distance between the backbone C^{α} atoms [7]. It can also be described by non-covalent side chain interactions derived from a distance based criteria [4], while in the case of an energy based network, edges are described using interaction energies between residue pairs [8]. The latter has the advantage of combining both the aspects of topology and the chemistry of interactions, thus making it more rigorous. Such an energy-weighted network defines not only the interactions between nodes but also provides information on the strength of interaction. Another way of defining an edge is on the basis of surface complementarities and the overlap between the sidechain surfaces in proteins [9]. Additionally, networks have also been built using criterions such as covariance data [10], evolutionary conservation [11] and amino acid fluctuations [3a] for identification of residues important for protein function, ligand binding and structural stability. Furthermore, connections in a protein structure network may be weighted or binary. Weights can be determined from the extent of geometric overlap or energetics of interactions. It can also be weighed on the basis of dynamic correlations between residues [12] or result from knowledge based pair potentials (for example, MJ potential [13]).

1.2.3. Conformational Dynamics in Proteins: Dynamically Weighted Structure Networks

In line with Richard Feynman's famous statement regarding the significance of the 'jiggling and wiggling of atoms [14], the study of protein structure-function is incomplete without an insight into conformational dynamics. Proteins or their complexes are highly dynamic undergoing significant fluctuations in their atomic level interactions. These fluctuations often code for important aspects of their function, often comparable to the more established influence from sequence and structure [15]. Consequently, an additional layer of complexity often has to be included to protein structure networks by allowing for the inherent conformation dynamics. This translates to the fact that not all edges in the network are stable during the protein conformational dynamics nor are they equally important. Such distinction can be made by weighing the edges, thus taking into account the motional dynamics. A dynamically stable edge weight can be computed by counting the occurrence of a particular edge in an ensemble of conformations. This provides a means to statistically evaluate the importance of a given edge in a network. So an edge may not only be weighed using energetics or correlation information but also by its statistical significance in the population ensemble or both. However, the criteria for selection of the nodes / edges, edge-weights or whether to include the dynamics in the computation of the structure network is determined by user discretion, driven mainly by the biological questions at hand. Table **1** provides a list of various types of protein structure networks and the criterion used to define the nodes and edges in the respective network.

2. TECHNICAL DETAILS

This section describes the network related mathematical and computational details of protein structure networks, in the context of structure and function. Section 2.1 describes various methods of network construction and Section 2.2 deals with some of the network parameters employed in elucidating the structural and functional aspects of proteins and their complexes.

2.1. Construction of Protein Structure Networks

Mathematically, a graph G=G(V,E) is composed of a set of vertices and edges, wherein two vertices V_i and V_j are connected by an edge E_{ij} if they are related. Depicting proteins using the mathematical formalism of graph theory [16] allows to efficiently capture the structural details of a protein for various analysis. Most of these networks are coarse-grained and generated at the level of C^{α}/backbone. However, the role of sidechain atoms in determining the structure and function of proteins is crucial. In our lab the importance of sidechain atoms are taken into account by constructing protein structure networks with emphasis on interactions between the sidechain atoms. In the subsequent paragraphs, the construction of three types of macromolecular interaction networks, namely, protein sidechain network, protein-DNA bipartite network and protein energy network are described. A graphical representation of the same is also provided in Fig. (**1**).

Table 1. Definitions used to construct protein structure networks.

Network Type	Node Definition	Edge Definition	Reference
Protein secondary structure networks	β – strands and α – helices	Sequential and spatial neighbourhood	[57, 87]
Protein Contact Network	Cα/Cβ atoms	Inter-atomic distance	[2a, 7]
Protein Structure network	Residue	Atomic fluctuations / knowledge based pair potentials	[3]
Protein Structure network	Any atom	Interatomic distance	[26e, 88]
Protein Structure network	Residues	Evolutionary conservation	[11]
Protein Structure network	Residues	MJ potentials [13]	[88]
Protein Sidechain network	Sidechain atoms	Interaction strength	[4]
Protein Sidechain network	Sidechain atoms	Interatomic distances	[89]
Hydrogen bond network	Sidechain atoms	Hydrogen bonds	[90]
Protein Energy network	Residues	Interaction energy	[8]
Protein Energy network	Residues	Cross-correlation	[12]
Protein Energy network	Residues	Surface complimentarity	[9]
Protein-DNA bipartite network	Two types of nodes, representing residues from protein and DNA	Inter-atomic distances between the protein and DNA, followed by a user-defined interaction strength cut off	[18]

2.1.1. Protein Sidechain Network

Protein Sidechain Network (PScN, also denoted as PSN in literature) efficiently portrays the non-covalent side-chain interactions in a protein from a global perspective. Amino acid sidechains are known to play an important role in defining protein structure and function. The details of the construction of such a graph at a particular interaction cut-off (I_{min}) and the implications of such graphs have been previously discussed in details [4, 17]. Broadly, protein sidechain networks are constructed by considering amino acid residues as nodes and edges are drawn between pair of nodes if at least one of their sidechain atoms pairs is within a cutoff distance of 4.5Å. Distance calculations are performed between sidechain atoms of sequentially non-neighbor residues (i-j≥2), except for glycine where Cα is considered for the analysis (Fig. **1a**). Each of the edges is assigned interaction strength (I_{ij}) as evaluated from Equation 1 (details given in reference 4).

$$_{i} = \frac{n_i}{\sqrt{\ _{i} \times}} \times 100 \qquad (1)$$

where, I_{ij}=Interaction strength between residue 'i' and 'j'; n_{ij}=number of atom pairs within the cutoff distance; $N_{i/j}$=Normalization value for residue 'i' and 'j', calculated using a large dataset of proteins to account for variable residue sizes.

The above equation estimates the strength of non-covalent interaction between any two residues, based on which protein sidechain networks are generated. Those pairs of amino acids that have interaction strength (I_{ij}) greater than a user-defined cut-off (I_{min}) are considered as connected in the network to give a protein sidechain network (PScN) corresponding to a given interaction strength cut-off, I_{min}. This criterion yields a binary matrix, called the adjacency matrix, representing structural connectivity. Generally, I_{min} values in the PScNs vary from 1% to 15%. The lower the I_{min}, the higher is the connectivity and vice-versa. Fig. (**1b**) provides a detailed graphical representation of the steps involved in construction of a PScN.

2.1.2. Protein-DNA Bipartite Network

The PScN described above is applicable not only to the monomeric structures of proteins, but also for multimeric protein complexes. However an extension of this formalism to macromolecular complexes like Protein-DNA complex is not straightforward. This is due to the fact that the edges are normalized with respect to the maximum number of contacts made by amino acids, as seen from a large number of protein structures. The normalization values for amino acids in proteins turn out to be within a narrow range due to the formation of a core. Such an advantage is not available for macromolecular complexes like Protein-DNA since a core may not be present at the interface and more importantly, the chemical nature of the basic units of nucleic acids are different from that of proteins. Nevertheless, a Protein-DNA/RNA complex can be treated as a bipartite graph [18] to characterize the macromolecular interfaces.

Bipartite networks contain two types of nodes, each type representing protein residues and DNA residues. Furthermore, a connection can be drawn only between nodes belonging to different categories. As a result, a protein-DNA bipartite network represents interactions that occur between protein and DNA. Similar to PScNs, non-covalent interactions between the amino acid sidechains and the nucleotides

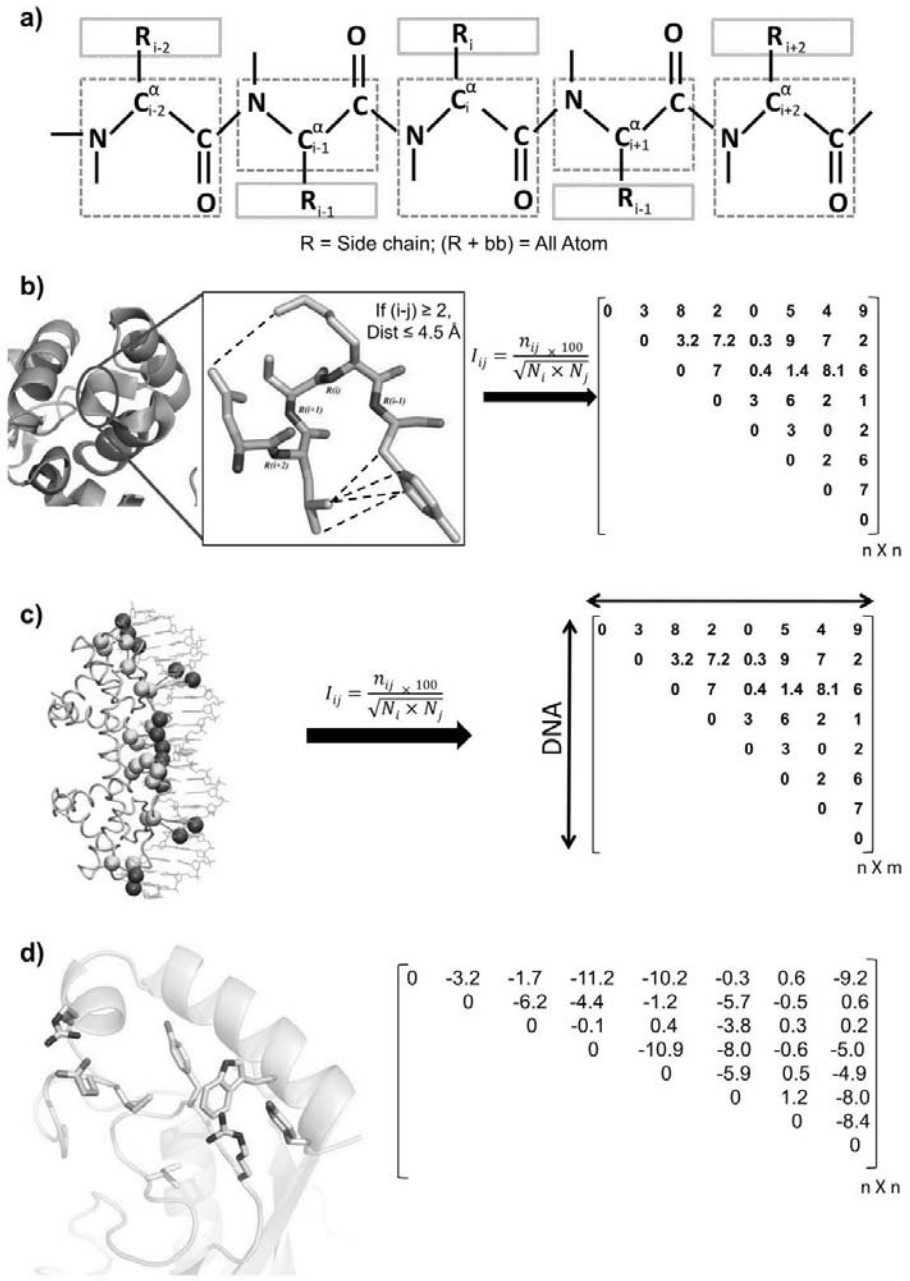

Fig. (1). Figure summarizes the methodologies used to generate the three types of Protein Structure Networks (PSN). (**a**) Describes the arrangement of sidechain atoms and backbone atoms in a protein sequence. While some studies use all atoms (except hydrogen) for generation of the protein structure network, protein sidechain network (PScN) uses only sidechain atoms for calculating interactions between residues. Interactions are calculated between sequentially non-neighbour residues, such that i-j≥2. (**b**) Graphical representation of PScN. Interaction strengths are calculated between residue pairs using the equation provided. This result in the generation of a matrix with size n×n, where n is the number of residues in the protein. Finally, based on a user-defined cut-off (I_{min}), an adjacency matrix describing the protein structure can be generated. (**c**) Protein-DNA bipartite network is a special case of PScN, wherein interactions are evaluated only between a protein residue and DNA residue. This leads to the generation of a bipartite network of size n×m, where 'n' is the size of the protein and 'm' is the size of the DNA. Adjacency matrices are then generated at a user-defined cut-off (MEC). (**d**) Pictorial depiction of PEN. The backbone of the protein (PDB_id: 2esk) is represented as transparent cartoon with a set of chosen residues shown as sticks. The interaction energy among them is represented as an 8×8 toy model PEN. The numbers in the matrix represent a range of interactions energies.

form the basis for linking the nodes. The strength of interaction (I_{ij}) between the amino acid sidechain 'i' and the nucleotide 'j' is evaluated in a manner similar to that explained in Equation 1. Based on a user-defined cut-off, termed as the Minimal Effective Connection (MEC), edges are drawn between the nodes. MEC quantifies the minimum number of atomic contacts expected between an amino acid and a nucleotide and can range from 0% to 15%, representing weak to strong interactions. The method is also represented graphically in Fig. (**1c**).

Three types of Protein-DNA bipartite network, representing the three fundamental units of a nucleotide (phosphate, base and deoxyribose) are generated separately to account for the different chemical nature of the subunits. Moreover, it is observed that the different entities of DNA can show different strengths of interactions. For instance, phosphate backbone of DNA mainly involves non-specific electrostatic interactions, while DNA bases are involved in specific non-covalent interactions. As a result, contacts are specifically evaluated between sidechains of the amino acids and the phosphate (PP), or the deoxyribose sugar (PS) or the base of nucleotides (PB), independently in order to capture the details of interactions of different fragments of nucleotides. Such a treatment has been found to be highly useful in characterizing Protein-DNA complexes based purely on atomic coordinates. For instance, bipartite networks have been used to characterize the nucleosome that stabilizes primarily by diffused electrostatic interactions with phosphate units whereas a predominance of base and deoxy-ribose interactions with amino acid residues is observed in beta-sheet proteins [18]. It has also aided to follow up the protein (IdeR)-DNA dissociation process, during simulations [19].

2.1.3. Construction of Protein Energy Network (PEN)/ Protein Complex Energy Network (PcEN)

The Protein Sidechain Networks (PScN) are constructed using geometrical features of the protein structure and hence can efficiently capture the topology and associated properties at the level of sidechain atom–atom contact. Although this is the simplest way to construct a PScN, the details captured by these network representations is incomplete and a wealth of information can be extracted by incorporating the details of chemical interactions. As advancement over the PSNs, PEN/PcEN [8] are generated wherein the edges are defined on the basis of pairwise interaction energies among the amino acids. PENs are energy based networks for proteins whereas PcENs are the same in the context of protein complexes with RNA/DNA. The interaction energy is the resultant of various types of interactions between a pair of residues within a protein (or between protein and nucleic acid) and captures all the essential features responsible for maintaining the protein structure. A pictorial representation of PEN construction is shown in Fig. (**1d**).

In energy-weighted networks, edges are constructed between all pairs of nodes (excluding the sequence neighbors) if interaction energies (E_{ij})>1 kcal/mol. PEN/PcENs are weighted networks with the normalized weight of an edge between nodes 'i' and 'j' (W_{ij}) defined as:

$$W_{ij} = -0.1|E_{ij}| + k \quad (2)$$

where, E_{ij} is the interaction energy between 'i' and 'j' and 'k' is a constant scaling factor. The incorporation of energetic terms in weighing the network not only makes it more quantitative, but also avoids the normalization values. Thus, the construction of energy-based network for macro-molecular complexes (PcEN) like protein-DNA complexes, though expensive, is more straightforward and efficient.

The interaction energy between two residues represents the edge strength and is also a measure of the ease of information transfer between them. E_{ij} can be evaluated by standard molecular simulations packages such as GROMACS [20], AMBER [21] and CHARMM [22]. The edge values can also be transformed to represent the cost of communication between any two residues. Therefore, the edge weights are normalized to represent cost by a simple linear transformation, so that the higher interaction energy connection (higher E_{ij}) is read as the lowest cost (lower W_{ij}). More the cost, less is the information transfer and vice versa. In our laboratory, the interaction energies between any two residues have been estimated using all-atom molecular dynamics simulations. However, other well-established knowledge based statistical pair potential energies (e.g., MJ potential [13]) can also be used to determine the interaction energy between residue pairs. This approach has the advantage of assigning weights without explicit all-atom calculation, but sacrifices the details of interaction, which may be unique to a specific case. The choice of method nevertheless is dictated by the problem at hand and computing efficiency.

Similar to the I_{min} based binary adjacency matrices of PScN, one can also obtain E_{min} based adjacency matrices at a given energy (E_{ij}>E_{min}) cut-off from PEN, such that all the edges with interaction energy greater than the cut-off are considered as connections. Thus, an energy-weighted matrix may be converted into a binary matrix at a chosen energy cut-off or probed using weighted network parameters, retaining the weights on the edges.

2.1.4. Dynamically Stable Protein Structure Networks: A Population Perspective

Whether the protein structures are translated into PSN or PEN, such a mathematical representation can either be employed on a single structure or on a structural ensemble. PSN/PENs can be computed for a crystal structure or over an ensemble of conformations obtained from MD simulations/NMR, etc. The PSN/PEN corresponding to a single crystal structure represents the interactions leading to a three dimensional organization in proteins or its complexes. However, when multiple structures are involved, two questions arise. The first question deals with the connectivity (in terms of either topology based or energy based side chain interactions) in a structural snapshot within the population ensemble. The second question pertains to the relevance of each of these interactions in the context of the ensemble. A user-defined cut-off is used to filter interactions that are present above a certain fraction of the population ensemble. This leads to the construction of dynamically stable PSN/PENs [8a, 23], which represents interactions within a population ensemble in a statistically relevant manner.

In the context of PEN, one can either generate PENs for every snapshot generated using MD simulation or an average PEN by averaging the E_{ij} values from individual snapshots as also shown in Equation 3:

$$_{i} = -\frac{1}{}\sum_{i=1} \; {}^{t}_{i}(|i - \;| > 1) \quad (3)$$

where, E^{t}_{ij} is the interaction energy between 'i' and 'j' in the t^{th} snapshot and 'f' is the total number of snapshots involved. This captures the average pairwise interaction energies from the MD ensemble.

2.2. Network Parameters: Exploring the Protein Architecture

One of the biggest advantages of network biology is that it enables the characterization, analysis and comparison of various aspects of networks using well-founded and mathematically derived network parameters. These network parameters can be used to characterize and analyze the PSNs/PENs/PcENs, which in turn can be correlated with different structural and functional properties of proteins. The parameters can be defined either at the level of individual residues or the network as a whole. Previous studies have used many such mathematical formalism to characterize a protein and hence identify hot spots [24], conserved residues [25], ligand binding residues and functional residues [26]. Many of the parameters are described in detail in literature. For the sake of completeness, some of the frequently used parameters have been described below and a graphical representation of the same is provided in Fig. (**2**).

2.2.1. Basic Network Parameters

a) Hub Nodes

Hubs are nodes that have a higher degree or connectivity (total number of edges incident on that specific node) in a network (Fig. **2a**). It can be evaluated by summing the number of edges incident on it or emanating from it. While in the case of undirected graphs, the two types of edges are identical, they represent different scenarios in a directed network. Unlike many other types of biological networks, the highest degree of a node in PSN is limited due to steric constraint. For the same reason, the highest degree of a node also depends on the strength of interaction and it could vary from three to ten in PSN. Thus generally, a node with a degree greater than three or five is defined as a hub in PSNs with higher and lower interaction strengths respectively. Similar definition also applies to nodes in PENs. Protein structures are known to be made up of a significant number of strongly and weakly interacting amino acid hubs so as to stabilize the tertiary structure of the protein. These hubs provide resilience against random mutations and form key residues important for protein structural stability and function [26a, 26c, 26e, 27].

b) Clustering Coefficient

Clustering coefficient provides a measure of the connectivity information about a node's neighbourhood [28], as opposed to hub that estimates the connectivity of the said node. Mathematically, the clustering coefficient of a node v, (C_v) measures the fraction of nodes connected to 'v' that are also connected to each other. C_v can be calculated using Equation 4.

$$C = \frac{2_e}{(\quad -1)} \quad (4)$$

where, C_v is the clustering coefficient of node 'v', k_v is the number of neighbours of 'v' and e_v is the number of connected pairs between all neighbours of 'v'. Clustering coefficient determines the influence of a node in a network. In general, any information that passes through a node with higher clustering coefficient would percolate faster in the network owing to the presence of a highly connected neighbourhood.

c) Assortativity Index

Assortativity index [2d, 29] of a node provides a measure of the average degree of the node's nearest neighbours. A positive correlation of assortativity represents the preference of a higher degree node to interact with other nodes that have higher degree, while a negative correlation is indicative of the tendency of a higher degree node to interact with nodes that have lower degree. Mathematically, it is calculated using Equation 5.

$$r_i = \frac{\sum_{=1}^{d(i)} d(\ )}{d(i)} \quad (5)$$

where, r_i is the assortativity index of node 'i', d(i/j) is the degree of node 'i/j'. It has been shown mathematically, that while social networks show high assortativity, other technology networks or biological networks (metabolic, signaling or gene regulatory network), tend to display lower assortativity. Assortativity has been discussed in terms of network resilience [29b], wherein it has been found that removing high-degree nodes in an assortatively mixed network would be an inefficient strategy for network destability, since it would result in destruction of members that are connected to each other and hence would destroy only a portion of the network and would be redundant. Interestingly, assortativity mixing is also considered to be one of the important factors for protein folding [30]. Protein folding is a cooperative phenomenon and requires communication between nodes. Additionally, assortatively mixed networks are known to have faster percolation of information, which is also a must for protein folding.

2.2.2. Parameters Associated with Percolation Behavior

a) Clusters

Clusters are a set of nodes in a network that are autonomously connected, such that the number of connections within the subset of nodes is higher than the number of connections these nodes make with the other nodes (Fig. **2d**). They are identified using the depth first search (DFS) algorithm [31]. DFS is a greedy algorithm that traverses deeper into the network. Due to the presence of higher connectivity amongst nodes in a cluster, these make a very good measure for predicting functional classes and novel annotations in a protein-protein interaction (PPI) network [32]. For instance, in a PPI network if an unannotated node, 'N' forms a part of cluster representing the functional class, 'x', then the function of N can be easily mapped to the functional class 'x' and further details can be extracted based on the other nodes it is connected to.

It should be noted that there is no unique solution to the cluster identification in large graphs. The identified subset may be a good cluster with high connectivity or may be a bad cluster with low connectivity within the subset. In PSN, good clusters are identified to detect domains and domain interfaces in multi-domain proteins [27, 33]. However, in the characterization of the network behavior of PSN, we look for the largest cluster (giant component as in the case of ER (Erdos-Renyi) network) [34] in the graph, which need not be a

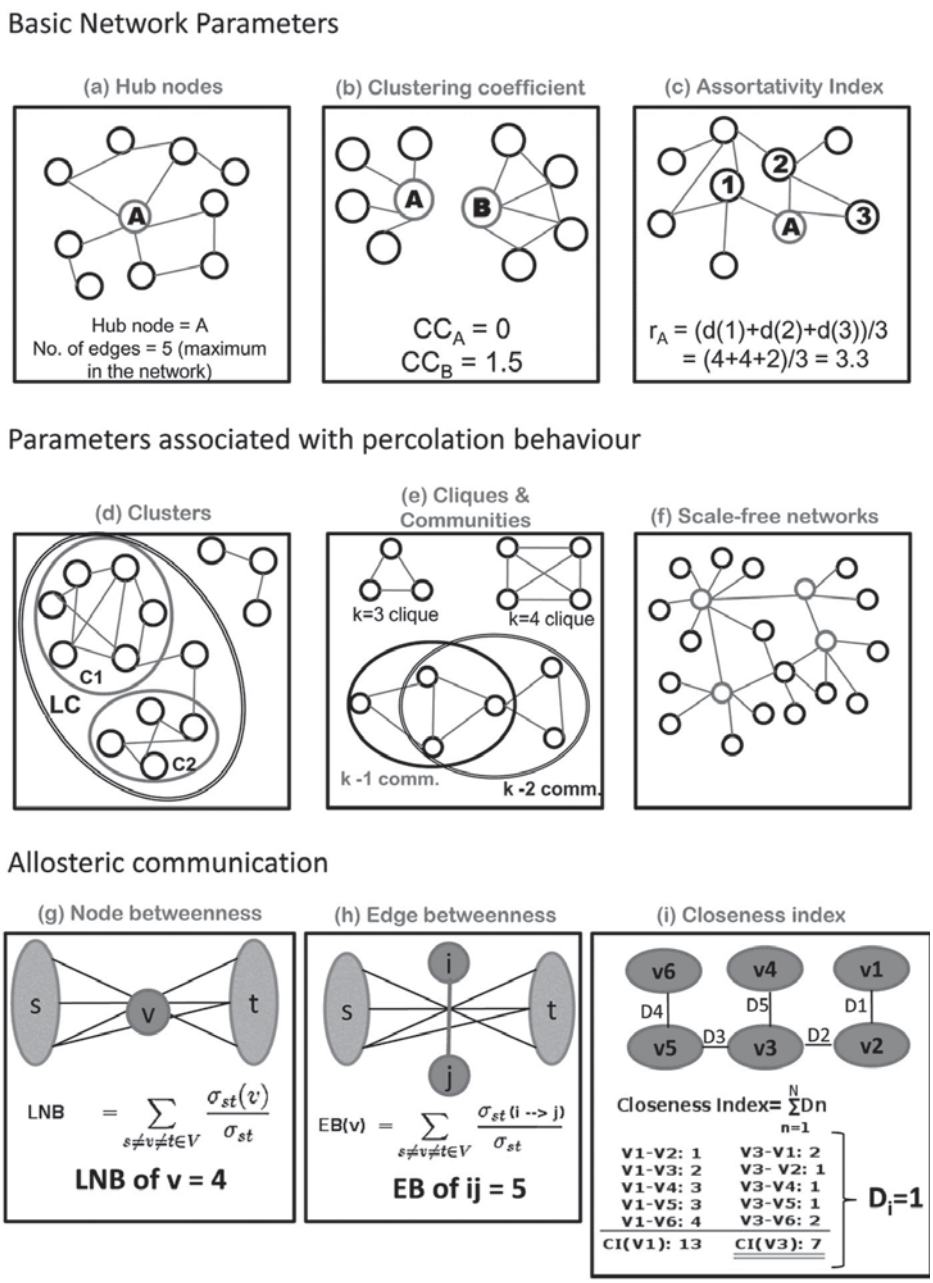

Fig. (2). Pictorial depiction of various network parameters. The parameters describing the properties of a node, such as (**a**) hubs, (**b**) clustering coefficient and (**c**) assortativity index. The parameters describing the properties of the whole network, (**d**) clusters, (LC represents the cluster with largest number of nodes, c1 and c2 are sub-clusters with high clustering coefficient), (**e**) cliques & communities and (**f**) scale-free behavior. The parameters related allosteric properties of the protein, (**g**) node betweenness, (**h**) edge betweeness and (**i**) closeness index. Details of each parameter are provided in the text.

good cluster from the point of view of intra-cluster connectivity. In a PSN, the largest component or the largest cluster defines the core of a protein structure and generally involves ~ 80% of all nodes in the network, when the edges are constructed at low interaction strength ($I_{ij} \leq 2\%$) and undergoes a transition to yield small disjointed clusters with medium interaction strength (I_{ij} between 2-5%). This is a common feature exhibited in protein structures in general [17b]. This also provides a novel method of assessing the quality of protein structures [35] and in studying percolation features in PSNs [36].

b) Cliques and Communities

A network can sometimes also contain complete subgraphs, such that each node in the sub-graph is connected to every other node in the sub-graph (Fig. **2e**). These are known as cliques and a collection of cliques are termed communities [37]. A k-clique is defined as a collection of k nodes, such that each node is connected to every other node. A simplest clique is formed by k=3 nodes. Collection of cliques that are connected to each other such that they share k-1 nodes are defined as k-1 communities, while if they share k-2 nodes,

they are termed as k-2 communities. Cliques and communities represent higher order connectivity in a PSN. They are used to identify the rigid regions in the protein structures and to recognize the ligand-induced conformational changes and presence of anchor residues in an active site or ligand binding site [23a, 38]. Communities on the other hand provide insights about percolation of the strongly connected components [36]. Percolation behaviour of protein structures as a function of interaction strength is also indicative of its quality [35, 39].

c) Scale-Free Behaviour (Comparison with Random Models)

Scale-free networks are characterized by a power-law degree distribution; the probability that a node has k links is given by $P(k) \sim k^{-\gamma}$, where γ is the degree exponent. The value of γ determines many properties of the system. For smaller values of γ, the role of the 'hubs', or highly connected nodes, in the network becomes more important. For $\gamma>3$, hubs are not relevant, while for $2<\gamma<3$ there is a hierarchy of hubs, with the most connected hub being in contact with a small fraction of all nodes [40]. Nodes in a scale-free network have a high degree of robustness against random failures, although they are sensitive to the failure of hubs. Previous studies in the laboratory [17b, 36] have compared global properties of protein structure networks to real world networks and random models. These studies have indicated that protein sidechain networks exhibit a complex behavior and do not behave as a scale-free network in a strict sense. However, they differ distinctly from random networks.

Here it is important to discuss the appropriateness of random model as null model for protein structure network comparison. The classical Erdos-Renyi (E-R) random model is not geometrically constrained and hence will be far from PSNs. Better null model would be to use a 3D constrained model as discussed in references [36, 41]. This is confirmed by a comparison of network profiles of PSNs with different types of random models, the classical E-R model (RM1), random model (RM2) constrained to protein size so as to fit the overall shape of a globular protein and nodes distributed randomly within a sphere of radius corresponding to that of the protein. In the third type of random model (RM3), the protein shape is strictly maintained, while the edges or connectivities present in a small radius (~6.5Å) of a chosen node are randomized. Thus RM3, which resembled the native protein structure networks very closely, is a better null model for comparison. Overall, it was observed that PScNs exhibit a complex behavior with combinations of Gaussian-like, sigmoidal, and exponential/power-law decay. More studies are required to clearly understand the unique complex behavior of PScNs.

d) PScN for Model Validation

As compared to the protein sequence space, which is astronomically large, the protein structure space is known to be much smaller. It is now understood that there exists ~1300 fold [42], with nature reusing these folds for different protein sequences. Examples of proteins having similar structure but varying protein sequences are abundant in nature. These selected set of folds have been optimized over time, such that they are rigid enough to provide structural stability and flexible enough to enable function. Given the constraint of a native protein structure, its comparison with a random model or decoys (computationally predicted protein structures) provides an effective measure for the identification of network features that are specific to protein structures. The uniqueness of the network behaviour observed in native structures prompted studies to use network features to identify native structures from a milieu of near-native decoys. This has been done using the machine learning (ML) technique, Support Vector Machine (SVM) [43]. SVM is trained to identify the native states through sidechain network features within the constraints of secondary structural elements. Here we describe the way the network features are being integrated into SVM.

SVM is a powerful discriminative supervised machine learning method that uses the task of pattern recognition for the purpose of classification. Intuitively, SVM uses a training dataset to learn the boundary between the two categories. The input data is mapped to a higher dimension, based on which a separating hyperplane is built, such that the distance between the hyperplane and the closest training sample is maximal. Given the complexity of biological systems, machine learning algorithms are widely used in biology to predict cellular locations [44], cancer tissue classifications based on gene expression data [45] and further in cases of protein structures to identify SCOP classes [46], binding sites [47] and also the quality of protein structures [48] using features, such as secondary structures and hydrophobicity.

The input data for SVM training in the context of PScN comes from the network parameters [35]. Adjacency matrices are constructed for I_{min}s ranging from 0% to 7% and network parameters such as NCov (number of edges), size of the largest cluster (SClu), size of the large communities (SLCom), clustering coefficient (CC) have been calculated at all I_{min} values. Table **2** provides a list of all the network features used for SVM classification. These network parameters, calculated at all I_{min}s (0% to 7%), along with the number of main chain hydrogen bonds (used for secondary structural constraint), resulted in a total of 94 features that were inputted into the SVM model for structural analysis. An important feature of the network behaviour is their characteristic transition profile, when plotted as a function of I_{min} (discussed in Section 3.2.2 and Fig. **8**). Transition of the profile is measured by calculating the difference in the parameter value obtained at consecutive interaction strengths (I_{min}). It is noted that the transition profile of a native protein structure differs significantly from that of the decoy structures, specifically between I_{min}=1%-4%. For training the SVM model, 3000 native and decoy structures were used. The training dataset was curated by randomly selecting structures from the master dataset of 5422 native and 29543 decoy structures [35].

The network parameters were further analysed to identify those features that truly capture the uniqueness of a protein structure. This study clearly established the importance of the transition profiles along with the main chain hydrogen bond for exhibiting significant difference between a native protein structure and decoy. On the other hand, the other network properties showed varied differentiating capacity at different I_{min} values. For instance, SClu, ComSk2, and CC

proved to be effective in capturing a native structure only at lower I_{min} values, while the k-1 communities could best capture the differences only at higher I_{min} values. While individually, the different network parameters capture the uniqueness of a native protein structure in different ranges, when the same parameters are integrated together into an SVM model, their effectiveness increases with an overall prediction accuracy of 94.3%.

Table 2. List of network parameters used for integration into SVM for the purpose of quality assessment.

Parameter	Description
NCov	Number of edges
SClu	Size of the largest cluster. Calculated using Depth First Search algorithm [31]
SLCom1	Size of the largest k-1 community, where k=3. Calculated using CFinder [37]
SLCom2	Size of the second largest k-1 community, where k=3. Calculated using CFinder [37]
SLCom3	Size of the third largest k-1 community, where k=3. Calculated using CFinder [37]
ComSk2	Size of the largest k-2 community, where k=3. Calculated using CFinder [37]
CC	Avg. clustering coefficient of the network. Calculated using algorithm described in this reference [91]
CC-LC	Avg. clustering coefficient of the largest cluster. Calculated using algorithm described in this reference [91]
CC-ComSk2	Avg. clustering coefficient of the largest k-2 community. Calculated using algorithm described in this reference [91]
d(NCov)	Transition profile calculated for NCov
d(SClu)	Transition profile calculated for SClu
d(ComSk2)	Transition profile calculated for ComSk2

2.2.3. Allosteric Communication

a) Paths of Communication

Allosteric communication between functional and allosteric/distant sites has a major role in the functioning of proteins and its complexes. It has been proposed to be mediated by a structured network of residues comprising the 'pathways of communication'. The 'structural view of allostery' posits that propagation of information across distant sites may involve interacting residues linking the distant sites. PSNs find application in determining paths of allosteric communication where well-established path-finding algorithms (like Dijkstra/Floyd Warshall) [49] can be easily applied to determine the shortest paths of communication and the key nodes/players/residues involved in the long-range signal transfer. A past review [2b] had addressed in detail some of the key developments and basic concepts.

Important conceptual and technical advances have been made lately to understand the nature of these communication pathways in proteins. These pathways, in the most simple of case, can be comprised of a linear 'daisy chain' of interacting residues/nodes [23b, 50]. However, a more complex network organization may be involved in long range signal transmission, involving interaction among collective rigid networks and/or residue-residue contacts [51]. These paths may emanate from a complex connection/interaction among conserved structurally rigid residues mediating global communication [52]. Important developments have occurred mainly on two fronts; (i) Defining new parameters to capture key residues which are not directly in contact with the ligands and (ii) Elucidating allostery from a dynamical perspective. The following sub-sections elaborate more on these points.

b) Weighted Networks and Paths of Communication: Concept of 'Junction Nodes'

In the recent years the literature has seen a massive upsurge in the methods of determination of key residues involved in allosteric communication [23b, 38, 53]. This has been in sync with the advancements in experimental techniques whereby more cases of allosteric regulation are coming to light [54]. Hence, along with analyzing the long range paths of communications, one can also study the possibility of information transfer through apparently non-interacting residues, which can be captured by network parameters. These parameters are different from the degree of a node as it involves the influence of the entire global network on individual nodes. It is usually more informative to compute these parameters from weighted networks (a detailed description of the various weighing methods has been described above). For example, centrality measures like closeness index have found application in determining key residues in the network in terms of its accessibility to information in the network. On the other hand, node and edge betweenness enable identification of residues and edges that are critical in information transfer in the network. In addition, parameters like 'junction nodes' have also been explored. These parameters are pictorially described in Fig. (**2f-h**). The 'Junction nodes' are the residues/nodes that flux maximal number of paths of communication between two chosen terminal residues or domains and hence form 'junction' points for signal transfer. This is similar to the concept of node betweenness with the flow of information restricted to user defined domains. These concepts have been applied to infer the key residues participating in allosteric communication between the active site and the tRNA binding domain in Pyrrolysyl-tRNAsynthetase [8a].

c) Pre-Existing Paths of Communications

Recent years have witnessed a paradigm shift in the understanding of allostery. Unlike the classical definition of allostery that was based on static structures, the newer definition is based on the conformational ensemble as represented by the shift in the energy landscape of the protein, with allosteric modifications defined primarily at the molecular level. The newer definition highlights the importance of dynamics for protein function, and hence, the dynamical properties such as equilibrium conformations and correlated movements of residues become important factors that require

exploration to understand allostery in a given protein. Some of these issues have been addressed in the literature. For example, the involvement of multiple preexisting pathways of communication has been postulated [55]. The new definition of allostery recognizes that the native states are ensembles of pre-existing populations and hence for a transmission of information there exist multiple paths, each of which are capable of long range communication. An allosteric effector leads to an equilibrium shift such that some of these pre-existing paths are activated while some remain nonfunctional. The studies on Pyrrolysyl-tRNAsynthetase [8a] have very clearly substantiated the concept of preexisting paths. The pre-existing paths have been very carefully quantified by calculating shortest paths for every snapshot of the MD simulations carried out for the system. The details for the same have been discussed in detail in a later section.

d) Sub-Optimal paths of Communication: Concept of 'Backup Pathways'

Similar to the concept of pre-existing paths, one can also define sub-optimal paths. In a dynamic ensemble the paths that are most frequently used to transmit signal are called the optimal paths. However, under abnormal cases, such as mutations or environmental changes, the proteins may reorganize itself such that another set of paths become more active in transmitting signal than the original paths. These set of new paths are known as sub-optimal paths. These are not the preferred paths of the protein under normal condition, but are preferred in the absence of optimal paths and hence termed as the 'sub-optimal' paths. The shortest paths (SPs) or optimal paths (OPs) of communication (lowest-cost path) between two residues are determined using Dijkstras algorithm. The cost of a SP between pairs of residues of interest is nothing but the summation of the edge weights constituting the path. Lower path cost implies higher efficiency of communication along that path and vice-versa. The suboptimal paths (SOPs) are the alternate routes of communication with costs greater than those of the OPs. Such sub-optimal paths have been characterized and quantified by systematically removing all interactions of an OP node(s), thus forcing the traversal of a less than optimal path in one of our studies involving MutS (manuscript in preparation). It should be noted that SOPs are computationally generated purely on the basis of the strength of communication paths within a protein in isolation. However, biologically they may also be considered as alternative paths, since the exact mechanism of communication in proteins embedded in the cell is not known.

3. PROTEIN STRUCTURE NETWORK: APPLICATIONS IN STRUCTURAL BIOLOGY

Network theory has found extensive applications in recognition of unique features in complex protein structures. The literature is replete with examples of the application of network theory in modelling various questions pertaining to the three dimensional organization of proteins and their functions. To name a few, the prediction of the catalytic sites or identification of critical residues participating in protein-protein/protein-nucleic acid interactions, atomic level structure networks has found extensive usage. Also coarse grain models like the elastic network models (ENM) and the anisotropic network models (ANM) [3a, 56] have been shown to capture large scale conformational changes. Several network theory based approaches have been developed to identify motifs in protein structure networks [26e, 57]. Further, spectral clustering of such protein structure networks has led to the identification of domains within proteins, which are important in regulation of protein function [33]. The coevolution of residues in protein towards a certain function in a correlated manner has been predicted by the analysis of protein sectors [58]. Reverse engineering techniques, to predict missing edges and links in networks, are also becoming increasingly popular in protein structure prediction algorithms [1b]. Understanding the link between the protein sequence and its fold has been a challenging problem. However, some advances have been made towards this goal. Characterization of fold-specific atomic interaction networks add to the collection of methods for protein structure predictions and model validation [11a, 59]. The properties of protein structures have been characterized through network features like largest cluster, cliques/communities [17b] which has enhanced our understanding of physical principles like percolation in a protein structure network [36, 60]. This unique percolation behaviour from a side chain perspective provides a metric to distinguish protein structure networks from random networks. This often becomes a useful tool to assess protein structures and *ab-initio* models from decoys. Furthermore, another key application of network theory is in evaluating critical players in allosteric/long range communication in proteins and its complexes with DNA and RNA [8a, 23b, 50, 61]. Only the last two applications mentioned here, namely, the allosteric communication in proteins and their complexes, and structure model validation from network approach are discussed further in this review.

3.1. Allosteric Communication: 'Action' at a Distance

The phenomenon of ligand binding at one site of a protein, inducing change at a distal site to elicit physiological response, is known as allostery in general terms. Ligands that bind to the first site are termed as the 'effector molecules' [62]. Classical models like the 'MWC model' [63] or the 'Pauling-KNF model' [64] have greatly advanced our understanding of the concepts of allostery or 'action at a distance'. Since, its inception many studies have been focused on explaining allostery in different proteins from various perspectives. While the basic dogma of the MWC model were the existence of symmetry related oligomers and the existence of discrete conformational states, the KNF model posits that the inter-conversion between these discrete states may occur in a sequential manner. A major conformational change was thought to be the key ingredient of mediating the long range communication among distal sites. Such a phenomenological view was supported by ground-breaking structural studies [65]. However, in the recent years, the literature has seen an upsurge in discussions about allosteric communication in proteins and its complexes. This has been stimulated by multiple experimental/theoretical observations triggered majorly by the technical advancements (both experimental techniques and computational power) in the last decade. With the literature brimming with examples of allosterically regulated proteins, there is rekindled interest to revisit the classical paradigms of allostery.

The exhibition of allosteric regulation in monomeric proteins [66], the discovery of multiple cryptic binding sites on proteins that are potent drug targets [67], the ability to engineer allosteric sites through environmental perturbation [68] have opened up new avenues to redefine our understanding of the mechanisms underlying allosteric regulation. More importantly from a structural perspective, the conformational changes can be as large as rigid-body motions to as subtle as side-chain reorientation [69]. Encompassing such a wide spectrum of examples begs for a generalized theory of allostery. The advent of advanced experimental as well as computational methods brought the ensemble view of conformations as the basis of this phenomenon. According to this view, the equilibrium structure of a protein is not a single structure but a collection of several conformations. This can be described in terms of the population of different conformations or in terms of free-energy landscape at a more fundamental level. Thus, the definition of allostery in terms of the shift in conformational population or energy-landscape due to perturbation (by the effector molecule or change in the environmental condition) is emerging as the basic principle of allostery [54b, 69, 70]. In fact, the same principle has been described to explain enzyme catalysis [71]. The ensemble view of allostery emphasizes on three things, a) change in conformational landscape as a function of effector binding (large scale as well as extremely subtle that eludes most classical analyses techniques), b) pre-existence of conformational states (population ensemble), albeit with globally unfavourable energetics, c) environmental factors such as pH, temperature and small ligands can also act as allosteric effectors, since these parameters have the capability to alter the population ensemble of a protein.

In light of this, it is also of utmost importance to pin down the underlying mechanisms of allosteric communication in atomic detail. Several experimental and computational approaches have been found useful, as has been discussed in detail in the past [8a, 12, 25b, 38, 51, 72]. While graph theory has found immense applications in exposing local connectivity patterns/information from a global perspective, MD simulations is an excellent tool to capture dynamicity in macromolecules [2b]. Here we have reported further advancements in our understanding of the principles of allostery, which invokes alterations at global structural level and redistribution of equilibrium populations.

Using the following examples, allosteric communication in light of its functional outcomes will be discussed. The effect of differential ligand binding has a major influence in reorganizing networks and driving signal transfer as will be elaborated below in the context of a protein-RNA/DNA complexes. In this review, a special emphasis is given to systems exhibiting subtle conformational variations accompanying the ligand binding.

3.1.1. Ligand Induced Rewiring of Protein Conformations

Upon ligand binding, proteins may exhibit drastic or subtle conformational reorientations. Such ligand induced changes often contribute to the functional outcome of the protein in a particular ligand bound state. Additionally, it is important to see the shift in conformation from an equilibrium ensemble perspective. Hence, it requires an integration of MD simulation with PSN analysis to understand such ligand induced changes in its entirety. Especially, subtle conformational changes that are mostly elusive to the established methods of structural comparison are efficiently captured by network theory in conjunction with MD simulations. Here we exemplify the advantages of this general methodology using Tryptophanyl-tRNAsynthetase (TrpRS) as a case study.

tRNA synthetases are pivotal enzymes ensuring the fidelity of translation of the genetic code. The prime reason for the choice of these enzymes for studies on allosteric communication emanates from their high efficiency long range communication between the active site and the cognate tRNA recognition/binding domain. Strikingly, such signal transfer is mediated by subtle conformational changes in most tRNAsynthetases, making them an ideal system for testing the efficacy of our network theory based methodologies. Specifically, in this first example, we emphasize on the identification of dynamic ensembles of proteins in its various ligand-bound states, followed by their characterization using network tools.

Tryptophanyl-tRNAsynthetase (TrpRS) is responsible for charging $tRNA^{Trp}$ with tryptophan with high degree of fidelity. The functional state of the human enzyme is a dimer with reported half-of-the-sites-reactivity. Ligand induced alteration of rigidity/flexibility and reorganization of the conformational ensembles, using concepts of dynamics and network theory has been reported for TrpRS [38]. The study involved the identification of the 'essential dynamical modes' from the PCA like analysis of the MD simulation trajectories [73] on various ligand-bound structures of TrpRS. A two dimensional 'essential plane' was constructed from the top two modes depicting more than 50% of dynamics and the simulated trajectories were mapped on to this plane to obtain an approximate estimate of the free energy peaks and their variations as a function of ligand binding [74]. By this procedure, free-energy landscapes have been derived for different liganded states. Although this is not a rigorous method for free-energy analysis, such a landscape is representative of the major conformational degrees of freedom and the conformational re-orchestrations in going from the apo state of TrpRS, to the fully ligand bound state. Furthermore, the integration with network analysis provides the subtle conformational changes due to ligand-binding, providing a microscopic look at the atomic level changes.

Deeper insights into the aminoacylation mechanism came from a systematic comparison of the low free-energy regions of the landscapes of the various ligand-bound states, in terms cliques/communities and hubs. The regions of rigidity and flexibility in the system are detected from these network parameters. Interesting aspects of the aminoacylation reaction became evident from these calculations. As the system transits from the apo to the Trp-AMP bound state, there is an increase in rigidity around the active site providing a rigid framework around the Trp-AMP. This rigidity gets slightly relaxed around the active site upon binding of the cognate tRNA-TRP, which may facilitate the penetration of the tRNA into the active site. It is noteworthy that although TrpRS is a homodimer, the subtle asymmetry between the two subunits becomes evident in the network analysis of the

conformational ensembles for the different ligand bound states. Such alterations in rigidity/flexibility in one subunit at a time, is in agreement with the concept of half-of-the-sites-reactivity. Such variations in rigidity/flexibility of different liganded states and induction of asymmetry in the dimer are also observed in another example of an a typical tRNAsynthetase called Pyrrolysyl-tRNAsynthetase (PylRS) [8a] (Fig. **3**) which was investigated by energy-weighted networks. Furthermore, ligand induced subtle changes in side-chain orientations, as captured by MD simulations and network parameters, have been characterized also in other tRNA synthetases like MetRS [23] and CysRS [50]. Based on these studies, a generalized model for the aminoacylation process in tRNAsynthetases has been proposed as shown in Fig. (**4**).

3.1.2. Energy-Weighted Structure Network: Insights into Long Range Signal Transfer in Protein Complexes with RNA/DNA

The information derived from coupling of network theory to MD simulations is further augmented by inclusion of quantitative chemical knowledge in the network. This is achieved by calculating pairwise interaction energies for all residues in the protein or its complexes with DNA/RNA and using these values to weigh an edge in the dynamically stable protein structure network (see Methods). These interaction energy weighted networks (PENs) are then probed to address questions of biological relevance. Ligand induced changes in rigidity/flexibility around the active sites have already been discussed above in the context of PylRS. Identification of key residues involved in communication, optimal and suboptimal paths of communications, and pre-existing paths are discussed below with examples of PylRS/RNA and MutS/DNA complexes.

a) Key Allosteric Residues and Alteration in Rigidity/ Flexibility

A specific application is described here in the context of allosteric communication in Pyrrolysyl-tRNAsynthetase (PylRS), which is responsible for charging $tRNA^{Pyl}$ with pyrrolysine in the absence of a precise tRNA anticodon recognition. Allosteric communication takes place between the tRNA binding site and the active site in PylRS. Crystal structure of the dimeric PylRS complexed with $tRNA^{Pyl}$ [75] reveals that unique interactions between the protein and the tRNA are mediated through a set of residues in the core binding surface of the protein and the key nucleotides called 'identity elements' on the tRNA. The critical residues participating in communication have been identified [8a] by probing the energy weighted networks derived from the MD snapshots of PylRS. Furthermore, ligand induced perturbations in the path and cost of communication have also been reported. These calculations supported the concept that function in PylRS was beyond its sequence-structure alone and was dictated partially by the dynamics of the interplay between sidechain interactions. These concepts are elaborated below with a comment on the generality of these ideas.

A detailed analysis of the networks derived from PylRS revealed the social architecture of the interaction network. Residues fluxing maximal number of shortest paths of communication between chosen modules were identified, which are thought to be critical for transmitting signal across long distances. Interestingly these 'junction nodes' (described in Section 2) undergo variations upon differential ligand binding and also from one subunit to another, highlighting the rewiring of these networks and the importance of capturing dynamical information in analyzing these systems. The nodes key to intermodule/interdomain communications are also identified by metric based on the concept of betweenness. In addition, global communication in such networks is further characterized through the parameter called funneling, which posits that there are few 'mediator' nodes that are central to communication throughout the network. In general, these metrics have potential applications in understanding the link between subtle conformational rewiring and allosteric communication in proteins. It is noteworthy that both from a local and global communication perspective,

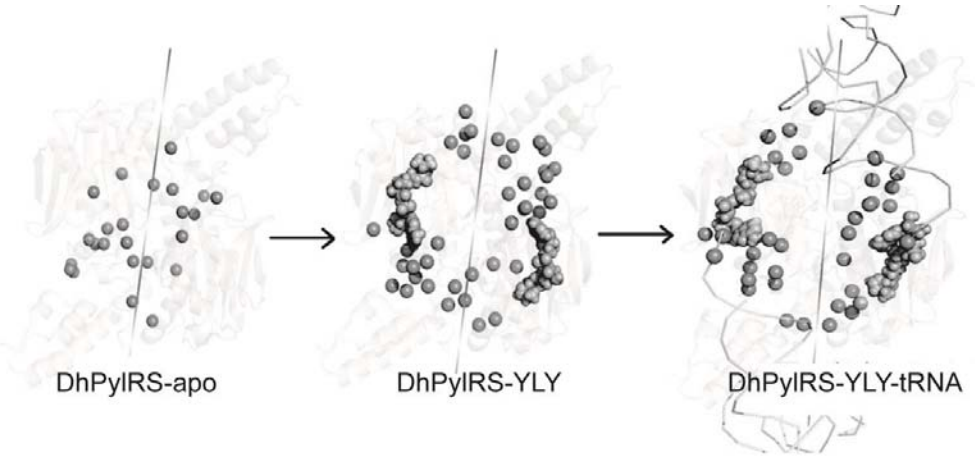

Fig. (3). Key residues in communication across functional modules (active site and tRNA binding domain in three different ligand bound states of pyrrolysyl tRNA synthetase. The C^{α} atoms of the key residues or 'junction nodes' are depicted as spheres. The ligand (YLY) is also shown as spheres. The protein backbone and tRNA are represented as a transparent cartoons. It is evident from the figure that there is a significant reorganization of the key residues upon different ligand binding. This emphasizes the subtle conformational changes as a function of ligand binding. Further, a marked asymmetry in the two subunits w.r.t. the critical residues is observed in the fully ligand bound state. (Reused with permission from [8a]).

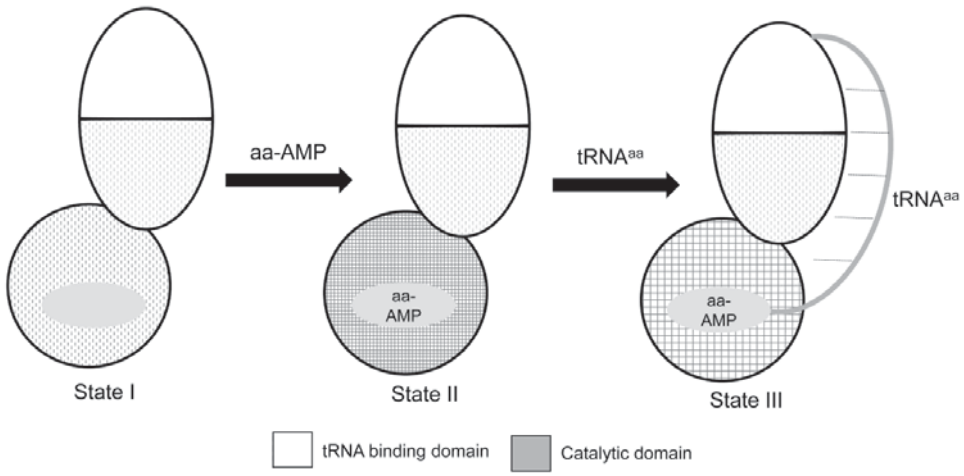

Fig. (4). General mechanistic scheme for the aminoacylation reaction. State I is the apo form, state II is protein bound to aa-AMP, and state III is the pretransfer complex (bound to aaAMP and tRNA). The grid density in a region of the structure is proportional to structural rigidity. A significant increase in rigidity around the ligand (aaAMP) is observed upon going from state I to state II, followed by a slight decrease in state III. The tRNA binding site also undergoes a slight increase in rigidity from state I to state II, which is dramatically increased in state III. This highlights the role of subtle conformational reorganization in terms of rigidity/flexibility in understanding the mechanism for aminoacylation reaction in general.

a marked asymmetry evolves during the dynamics between the two subunits of a structurally symmetric homodimer in the appropriately ligand bound state. However, such asymmetry is not observed in the apo state, relating asymmetry to function. The connection of ligand binding with the development of such asymmetry asserts the significance of ligand induced subtle variations in dynamics and sidechain interactions towards function, in the absence of major conformational alterations.

To summarize, the cost of long range signal transfer between the communicating modules illuminates upon the ease of information transmission between these sites. It is evident from the evaluation of the communication cost, in terms of energetics of interaction between the active site and the tRNA binding site, that appropriate ligand binding optimizes information transfer across these sites by suitable rewiring of interactions. What is most interesting, with specific reference to PylRS, is asymmetry in the efficacy of one subunit over the other in terms of signal transfer. This directly corresponds to the experimentally established observation of half-of-the-sites-reactivity in structural/interaction terms in tRNA synthetases in general [76].

b) 'Allosterism' and Pre-Existing Paths of Allosteric Communication

The conformational space of a macromolecule is constituted by an ensemble of conformational populations that are in dynamic equilibrium. In keeping with this concept, it is proposed that there is also an ensemble of pathways for allosteric communication [54b, 70b]. Although this concept is highlighted in the literature, a rigorous and quantitative evaluation of the same has not been reported. In this review, we have probed this hypothesis using a combination of network theory and MD simulations. The merging of network manipulation with dynamical information from MD simulations indeed facilitated a direct visualization of the multiple pre-existing pathways of allosteric communication in an ensemble of conformational populations. This is elaborated with examples PylRS/RNA and MutS/DNA complexes.

In addition to considering paths of communication through optimal pathways, the existence of suboptimal paths (SOP), which are energetically comparable to the optimal paths are also explored by PENs of the two systems. SOPs can be mathematically generated by recalculating OPs from deletion (node/edges) network. Deletion networks are generated by systematically deleting the nodes constituting the OPs. From equilibrium ensemble point of view, these SOPs can be compared with OPs of sub-populations in simulations or from conformational populations from different liganded states of the protein. It is interesting to note that a relationship does exist between the SOPs in different populations from simulations, as shown in Fig. (**5**). Analysis of such SOPs led to the general notion that states of system that are optimized for signal transfer evolve (whether by major conformational change or through subtle variation in dynamics of interaction) to maximize their chance of success with multiple degenerate pathways for communication. This also fits in with the broad evolutionary view of degeneracy driven robustness in function by providing multiple accessible routes to success for any biological function.

The optimal paths of communication, the changes observed in different liganded states and the shift in the population of different OPs are quantitatively analysed in the case of MutS and its complexes. The fidelity of DNA replication coupled to the efficacy of repair mechanisms ensures integrity and genome stability within a living cell. Mismatch repair system (MMR) scans DNA for any replications errors, leading to repair. MutS is a principal component of the MMR and forms the first check point for replicated DNA. MutS scans the DNA for base-base mismatch and insertions/deletions as well as damage lesions. Once recognized, MutS interacts with the mismatch using a highly conserved Phe-X-Glu motif through base-stacking and H-bonding as

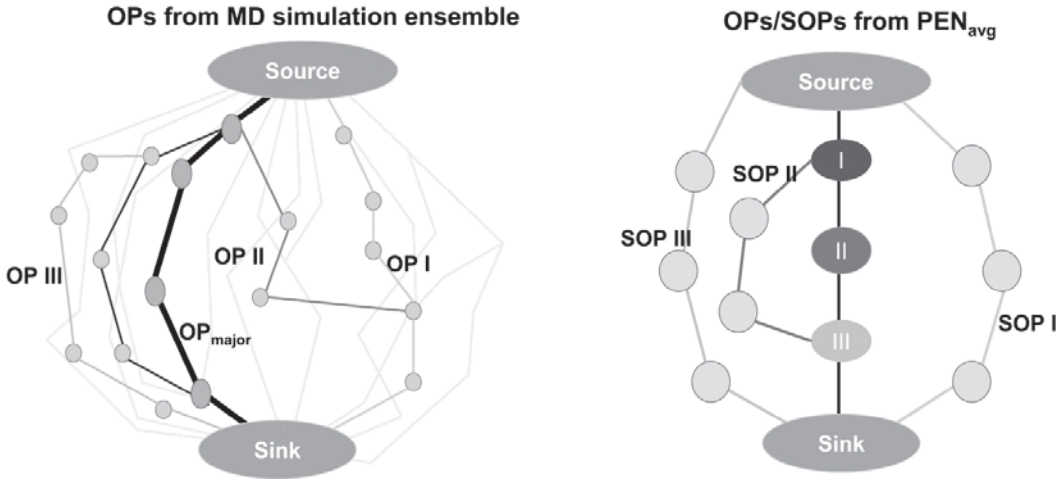

Fig. (5). Pictorial depiction of the unique sets of OP/SOPs between source and sink. (**a**) OPs from the major conformational population of the MD ensemble is represented as a black line and fine grey lines represent minor populations. (**b**) OPs and SOPs between source and sink from PcENavg. The black line represents the OP. Grey lines depict first-generation SOPs. Most of these SOPs appear as OPs in a small fraction of the MD ensemble as indicated by the corresponding labels (OPI/SOPI, etc). (Reused with permission from [8a]).

well as other non-specific interactions. This is the only base-specific interaction with the mismatch and the rest of the DNA-MutS interaction is highly non-specific to ensure a broad repair coverage. Allostery has been elucidated by analyzing the trajectories from extensive simulations on MutS and its complexes [77]. We have recently studied allosteric communication in MutS using the network theory tools on simulation trajectories. A detailed discussion of the results will be presented elsewhere (manuscript in preparation). However, for the purpose of this review, the aspect of pre-existing paths of allosteric communication is presented below.

Upon comparison of the shortest path/optimal path (OP) of a perturbed network with the original network derived from MD simulations (see Methods section), often the OP in the perturbed network pre-exist as an OP in a small conformational population in the original network. A particular example for a path between a pair of residues chosen from the ATPase and the DNA-binding module is shown in Fig. (**6**). Thirty five OPs are evaluated between the residue pair 14P-585N for all the MD snapshots considered for analysis. Among the thirty five OPs, the one that is present in the highest fraction of MD population is termed as dynamically stable. The dynamically stable OP is traversed by about 33% of the MD ensemble. This dynamically stable OP is blocked by systematic deletion of interactions of its constituent nodes (as explained in Section 2), thus generating the perturbed network. The OPs between 14P-585N are recalculated for all perturbed networks corresponding to all the snapshots. It is worth mentioning that the most populated alternate path or OP-perturbed actually pre-existed in the original network in a minor MD population. This observed feature, elaborated for a specific path, is seen for majority of the paths between the two important modules (the ATPase and the DNA-binding module) in MutS (data not shown). These results strengthen the theory for 'pre-existence of communication paths'. To the best of our knowledge, this is the first analysis to examine 'pre-existing pathways of allosteric communication' in a quantitative manner.

3.2. Model Validation and Network Theory: Differentiating 'Decoys' from 'Real'

3.2.1. Relevance of PSN for Model Validation

The central dogma in structural biology that 'protein sequence dictates the conformation' was established by the classic work of Christian Anfinsen on Ribonuclease in 1960s [78]. As a general rule this statement is very much valid. However, now with the availability of a large number of protein crystal structures there are stray examples where highly homologous sequences can adopt different conformations. More importantly, it is becoming increasingly clear that the number of protein folds are highly limited (of the order of ~1300, [42]) and non-homologous protein sequences have adopted the same fold. Here we would like to emphasize that the protein structure goes beyond one dimensional sequence. It is not just the sequence, but the interactions of side chains that are important in adopting a specific fold. This point was explored through PSN by capturing fold-specific interactions in Rossmann fold [11a]. Furthermore, the sidechain interactions in a protein results in a three dimensional network and the network architecture at different interaction strength seem to obey a unique behaviour as discussed above. These facts bring emphasis on two points. The uniqueness of protein structures is determined by (1) interaction among side chains and (2) a percolation behaviour in the three dimensional structure, which is distinct from random networks. These observations have provided an excellent ground to utilize the protein sidechain network for validating the protein models. The availability of a large number of native structures and their decoys in literature has enabled to train Support Vector Machine (SVM) to classify native structure from decoys. Thus any protein structure model can be evaluated for its native-like network behaviour.

Protein sidechain networks are sensitive to subtle alterations in sidechain conformation and hence serve as an attractive concept, providing useful tools for quality assessment of modelled structures. Using large scale data of native and

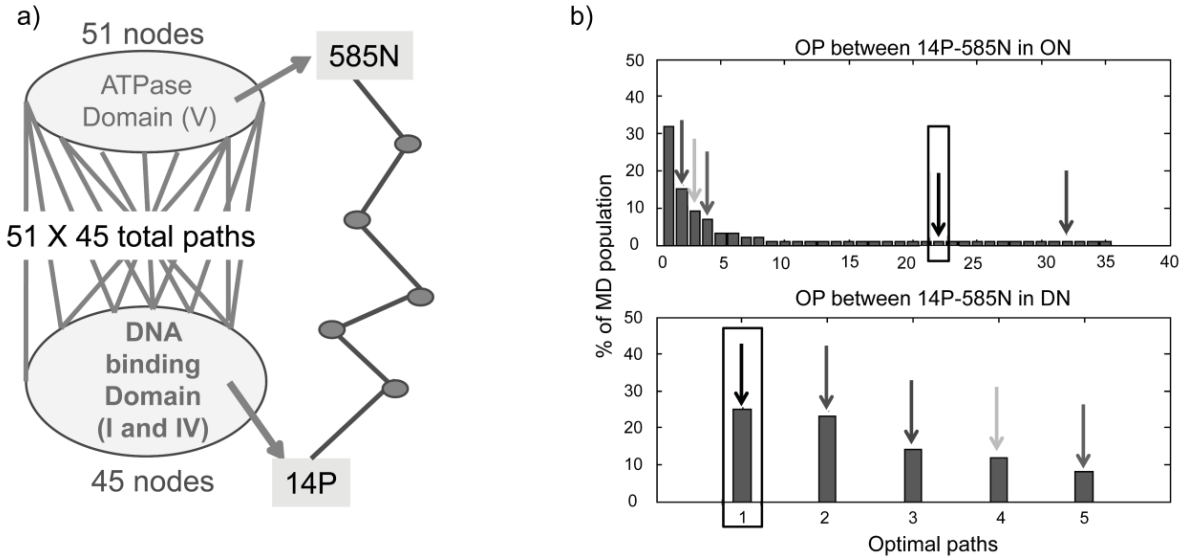

Fig. (6). (**a**) Schematic showing the evaluation of all optimal paths between the ATPase and the DNA-binding module, highlighting the path between an example pair (585N →14P) in MutS. (**b**) Thirty five OPs are evaluated between the residue pair 14P-585N for all the MD snapshots (top panel). The one that is present in the highest fraction of MD population is termed as dynamically stable and is traversed by ~33% of the MD ensemble in the original network (ON). Systematic deletion of interactions of its constituent nodes generates the deletion network (DN). The five sets of OPs in the DN and their representation in the MD ensemble are shown in the lower panel. It is worth mentioning that the most of these alternate paths or OPs in deletion network pre-existed in the original network in a minor fraction of the MD population (as shown by corresponding sets of arrows).

decoy structures, PSNs are explored to determine its capability to assess the quality of a predicted protein structure model. Development of a model validation protocol (Fig. **7**) enabled the characterization of the network features of native as well as decoy structures and specifically, these features served as guiding principles to determine ranking schemes for model evaluation. A large scale community-wide prediction of protein structures and evaluation of the predicted models have been facilitated by the Critical Assessment of Structure Prediction (CASP) group [48b, 79] and forms an excellent resource to obtain a set of such native and non-native-like structures.

Discrimination of natives from decoys was apparent through network parameters, even in the weakly interacting, highly connected regime (I_{min}=0-2 %), when carried out on 150 native and their decoy sets [39]. The parameters, clustering coefficient of the largest cluster (CClu) and the size of the largest community (SCom) were found to be highly sensitive in differentiating structures from both the categories. Interestingly, these are also the parameters that represent higher-order connectivity in a network, highlighting the importance of optimized sidechain packing in a native protein structure. Nevertheless, consideration of the complete network behaviour, including the percolation regime and strongly interacting, small nucleation clusters regime improved the performance to a great extent, supporting the idea that the native structures have unique percolation-like behaviour. This led to a rigorous development of PSN based quality assessment (PSN-QA) algorithm [35]. Details of training the dataset through the machine learning (ML) technique known as the Support Vector Machine (SVM) are provided in Section 2.2.2(d) and it's performance evaluation is presented in the following section.

3.2.2. PSN-QA Algorithm

A large dataset of 5422 native structures and 29543 decoy/modelled structures have been used to develop a mathematical tool based on PSN for quality assessment (PSN-QA). A major portion of these models were obtained from previous CASP experiments (CASP 3/7/8/9/10) thereby increasing the heterogeneity in model quality.

The selection of features is the crucial component of SVM. In the context of PSN, the features were selected to capture the network profile, in addition to the backbone constraints. The backbone constraint has been considered in a straightforward manner by evaluating the number of backbone hydrogen bonds. This accounts for the secondary structural constraints that are responsible for keeping the helices and sheets intact. However, more involved procedure is required to capture the features to reproduce the network behaviour. For each of the models studied, PSNs were generated at I_{min}=0% to 7%, followed by evaluation of nine network features. These features represent network parameters such as the total number of non-covalent interaction (NCov), sizes of the large clusters (SClu), communities (SCom), and clustering coefficient (CC) as presented in Table **2**. Interestingly, each of the nine network features evaluated for, exhibited similar profile for all native structures and had three characteristic features, 1) higher values at lower I_{min}, 2) lower values at higher I_{min} and finally 3) steep transition profile between I_{min} =1% - 4% when compared to the decoy structures (Fig. **8**). In order to capture the network profile as a function of interaction strength, the gradient of the curve is denoted by the difference in the size of the parameters as a function of I_{min}. More importantly, these three features together could distinctly differentiate between a native and a decoy structure. These native-structure specific properties

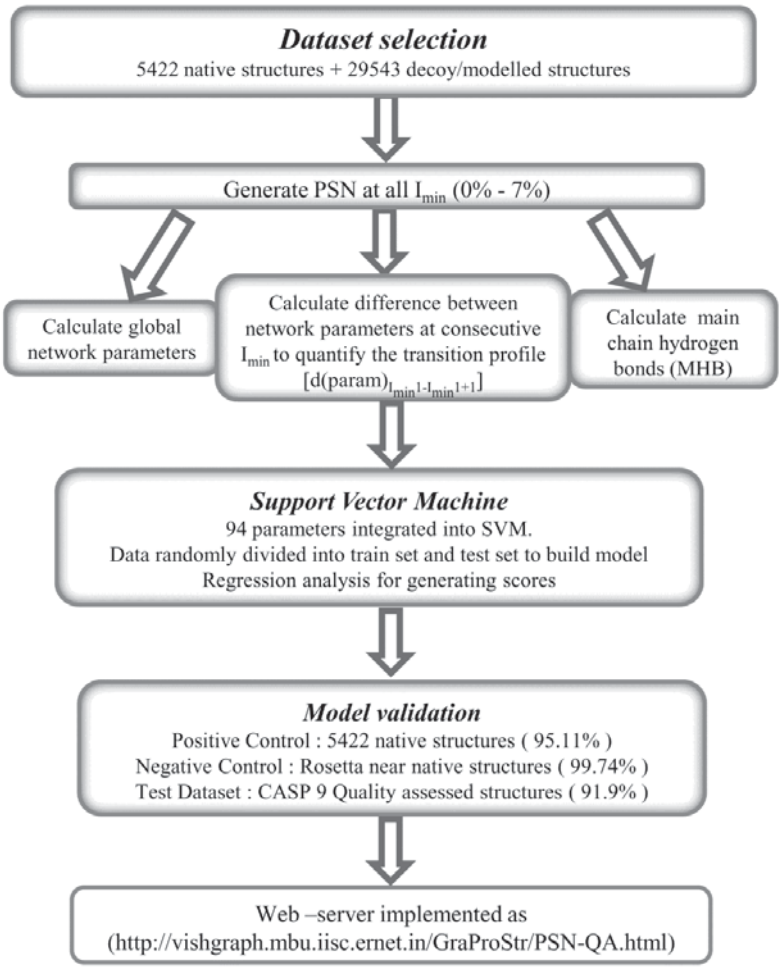

Fig. (7). A flow chart describing the steps involved in PSN-QA. Each protein is converted into PScNs, followed by calculations of network parameter. The parameters calculated at different interaction strengths are integrated into SVMs to generate a classifier model. Subsequently, regression analysis is performed that generates scores to estimate the quality of protein structure models and rank a set of decoy structures.

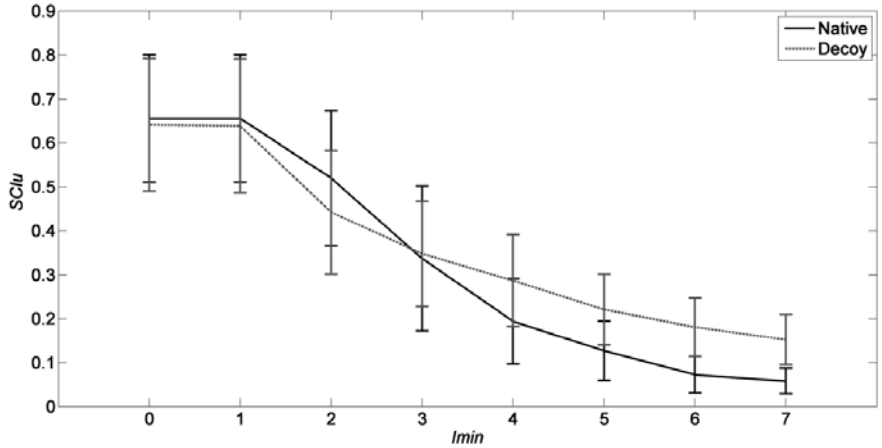

Fig. (8). The plot describes the profile of the size of the largest cluster (SClu) as a function of I_{min} for native (continuous line) and decoy (dashed line) structures. The characteristic transition profile is evident from the figure.

were then integrated with support vector machine (SVM) to obtain a mathematical framework for assessing the quality of predicted protein structures, as presented in Section 2.

The SVM method provided a classification accuracy of 94.22% [35]. Further, the quantitative estimate uses regression analysis to generate a scoring scheme for the candidate models [80]. The percentage score distribution of a set of native and decoy structures is plotted in Fig. (**9**), clearly suggesting that structures with score greater than 16 can be considered as good models, while those below 10 are considered

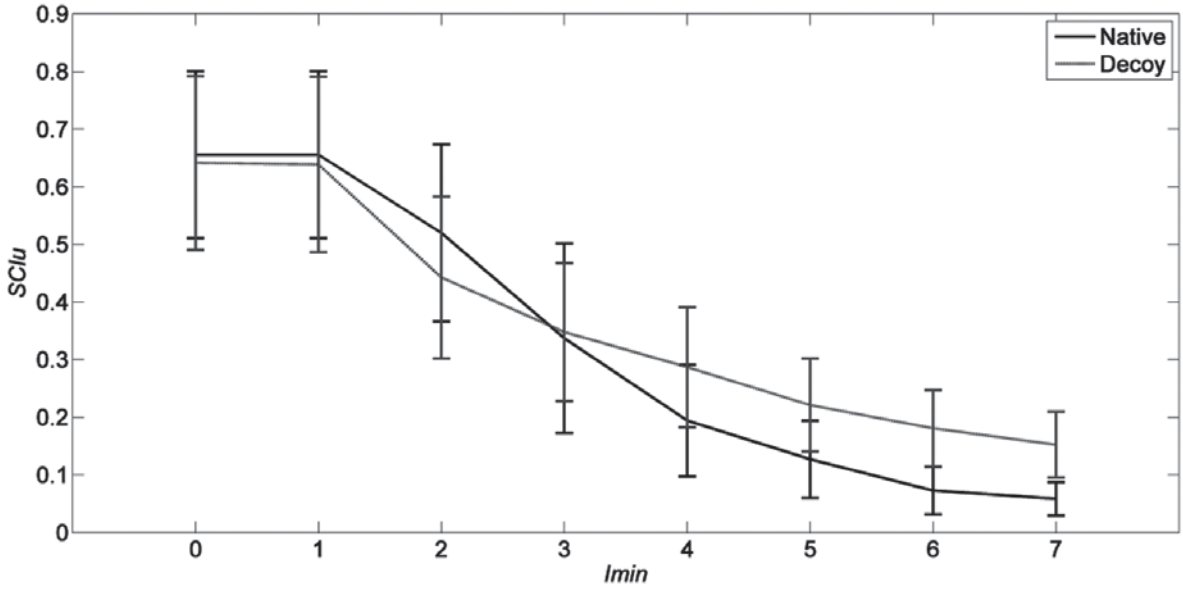

Fig. (9). Score distribution of the native and decoy protein structures as obtained from PSN-QA. Structures with score greater than 16 can be considered as good quality. These scores can further be used to rank a set of decoy structures.

as bad models. Few models lay in the transition region and have scores greater than 10 but less than 16. Additionally, the scores so obtained can also be normalized to lie between 0 and 1, by mapping the score of 16 to 0.5 and so on, as also illustrated in the figure. PSN-QA is the first rigorous attempt to make use of the global topological features that are sensitive to interactions between sidechain atoms to assess candidate models. It is based on the global structural features of the proteins and hence does not require any reference structure for quality estimates. The uniqueness of a native three dimensional protein structure is also very effectively captured by the native structure profiles. The study highlights the importance of analysing parameters at the backbone as well as the sidechain atom levels to predict model quality. In fact, PSN-QA is successfully able to identify structures that are well packed, have well-formed and properly oriented secondary structures. In the next section, we briefly describe examples cases, where PSN-QA has been successfully applied.

3.2.3. PSN-QA: Assessment and Applications

PSN-QA was successfully applied to assess the quality of CASP10 predictions [35]. A total of 31869 predicted models were assessed for their quality, followed by comparison with the GDT-TS scores as provided by the CASP organizers as a measure of model quality. Quality assessment was considered correct if the model was also identified as good quality according to the scores provided by CASP10. Of all the models identified as good models by PSN-QA, 71.4% were also selected as good models based on the GDT-TS scores, in the target-based modeling category. The percentage increased to 86.6% for the free-modeling category.

In the recently concluded CASP11, PSN-QA was used as an integral part of a collaborative pipeline, WeFold [81] for the model refinement category. The role of PSN-QA in the pipeline was to assess the quality of the refined models, based on which the top 5 structures were selected and submitted. CASP11 provided a total of 35 refinement targets, which were refined by the WeFold community. Of these 35 targets, PSN-QA was involved in identifying top 5 refines structures for 17 targets. In all cases, model refinement showed an increase in the PSN-QA score and the loop structure improvement by a PSN-QA refined model submitted from WeFold was noted by CASP11 assessors [81b].

3.2.4. Capabilities and Limitations of PSN-QA

The decoy models in literature and the models from CASP experiments have aided in evaluating the performance of PSN-QA. The major strength of this network based program is that it does not need a reference structure to compare. In other words, a rank can be assigned to any modeled/decoy structure. This has tremendous advantage of selecting a small subset of structures for further refinement from a vast pool of models generated by different model building programs. PSN-QA can help in drastically reducing the structure space for rigorous refinement in a computationally efficient manner. However, the quality assessment results from PSN-QA are applicable mainly for near-native structures. Applying PSN-QA on refined models based on bad structures can be a futile exercise. Hence a judicious choice of integrating PSN-QA at appropriate stage of the model building pipeline can enhance the structure prediction accuracy. The CASP experiments may also aid in improving the performance of PSN-QA.

4. SOFTWARES/ WEBTOOLS

4.1. PSN-Ensemble: A Comprehensive Tool for Dynamic Network Analysis

Network theory, as explained above for some specific cases, has found wide applications in understanding the physico-chemical details underlying complex biological principles. Therefore it is timely to build automated, robust

tools in order to expand its usage to a wider biological community. This has been made possible in the recent years with the advent of several software packages and webservers providing network based analysis tools for protein structure networks, like WISP [82], Xpyder [83], NetworkView [84]. Network analysis of single structures provides useful information of local interactions in the global milieu as is exemplified by GraProStr [85]. However, it is imperative to include the dynamic perspective for a comprehensive understanding of macromolecular function. Consequently, there is a pressing need to develop a consolidated and automated tool for analyzing conformational ensembles using concepts of network theory. This led to the development of PSN-Ensemble [86] (Fig. **10**) in our lab.

PSN-Ensemble is a robust, automated standalone tool for analyzing protein structural ensembles derived from MD simulations or NMR studies, from a network perspective. Specifically, network parameters such as hubs, clusters, cliques/communities, paths of allosteric communication between two residues, and important nodes for inter-domain signal transfer can be calculated using PSN-Ensemble for a conformational ensemble. The details of the program and its usage have been described in detail elsewhere. Here we summarize a few key points that are unique to this program:

1. Evaluation of sidechain interactions based network parameters at the desired level of non-covalent interactions.
2. Network features for conformational ensembles, derived from either MD simulations, NMR studies or using multiple crystal structures as input.
3. Allosteric communication related parameters are given eminence in view of their importance in understanding important biological principles.
4. Various weighing schemes for the network is provided, in terms of interaction energy or dynamic cross-correlation or ensemble weight, to provide user end flexibility.
5. Mapping back of all results onto the three dimensional structure in Pymol for publication ready depictions of the network parameters.

These key features of PSN-Ensemble has been exemplified with several examples, ranging from MD conformational ensembles of an ubiquitin ligase (UbcH5b) to crystal structures of different ligand bound forms for β2-adrenergic receptors and tyrosyl tRNAsynthetase. An array of questions pertaining to important biological phenomena, from allosteric communication to ligand induced optimization in structure network, has been addressed.

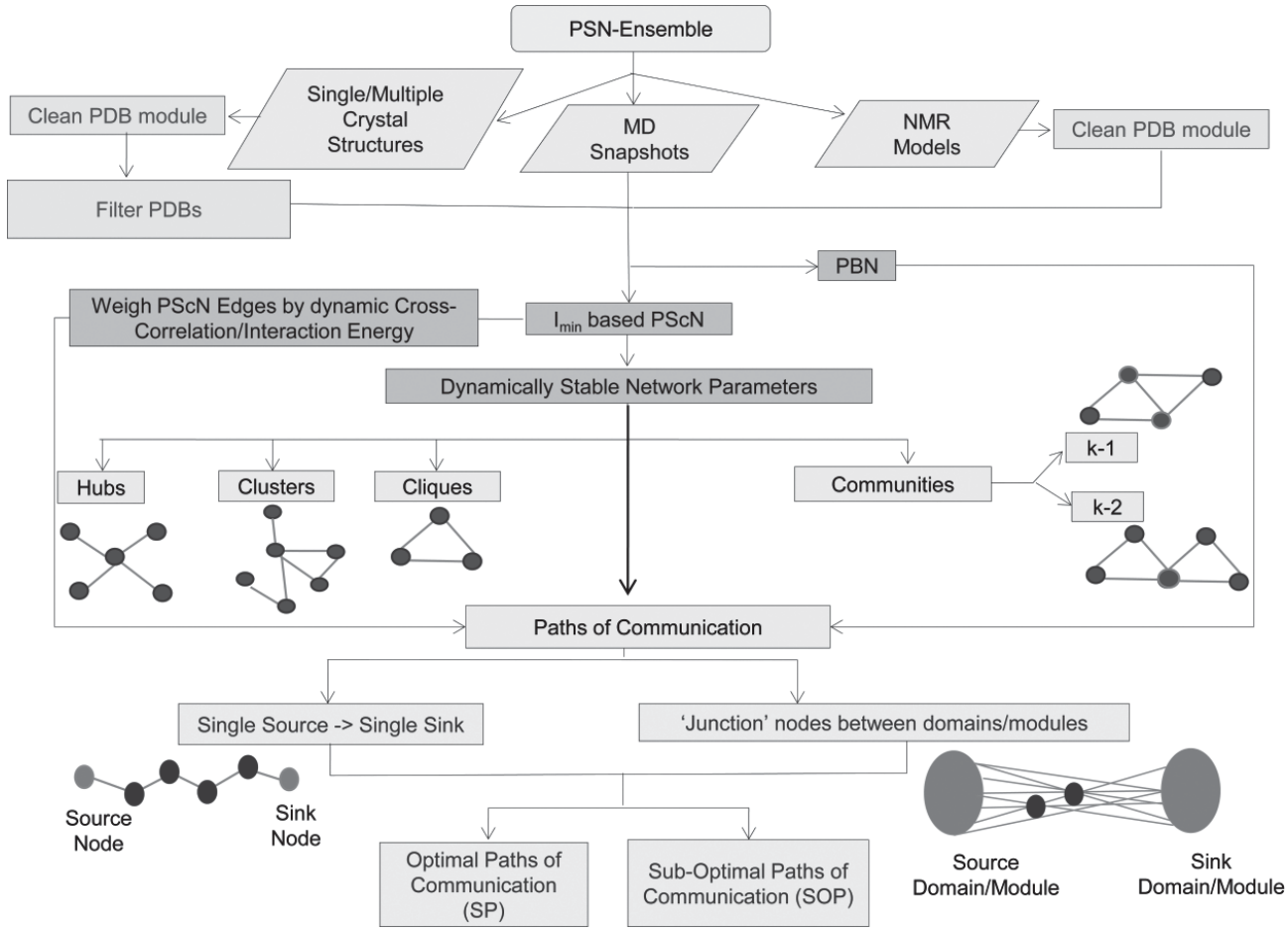

Fig. (10). Flowchart representation of the workflow in PSN-Ensemble. Various network parameters can be computed using this standalone package for crystal structures (single or multiple structures), MD ensemble, and NMR structures. Schematic description of each of network parameter used is shown in the flowchart for clarity. (Reproduced with permission from [86]).

4.2. PSN-QA: A Web-Server to Asses and Rank the Quality of Models

PSN-QA [35], a tool to predict the quality of protein models and rank them is also made freely available for public use in the form of a web-server, http://vishgraph.mbu.iisc.ernet.in/GraProStr/PSN-QA.html. The method is presented as an independent module in the webserver GraProStr [85]. The tool can take up five structures at a time and provide an estimate of the quality of protein models. For structures with multiple chains, individual chains are treated as different structures for the analysis. It accepts files in PDB formats as input and outputs the quality and rank of each model in a tabular format. A test case (PDB Id: 1CG5 and its decoy structures from Rosetta) is also provided with its scores as an example. Additionally, the files containing the network parameter values at different I_{min}s and the plots showing the transition profile are also available for download.

5. FUTURE DIRECTIONS AND PERSPECTIVE

Enormous body of experimental and theoretical work has been carried out to crack the code for protein folding and understand their function at molecular level. Yet there is a need for new outlooks and methods to fully understand the rules governing these processes. One can ask if the network approach has anything more to contribute to this problem. The problem can be addressed at different levels. For example, in the present review we have elaborated on principles of allosteric communication and discussed some preliminary applications towards the model validation problem. (1) The communication within and across proteins can further be explored through multi-scale approach by combining network methods with other studies such as simulations, mutational correlations and protein-protein interactions to understand molecular details of protein-protein interaction in cells. (2) Can we gain insights on protein folding from the vast amount of available protein structural and sequential data? Several attempts have been made in the past to identify the fold specific signature motifs. In this context, sequences with remote homology adopting the same fold could become an excellent source of data for extracting network based rules. (3) The fundamental process of protein folding may be simulated and properties such as percolation can be investigated at structure network level, to understand the physical principles of this phenomenon. It has been shown in the past that proteins exhibit unique percolation properties which can be exploited to validate models. (4) Mathematical tools such as network perturbation can be adopted to spot the influential amino acid nodes in the networks and correlate the results with diverse available experimental data such as mutations in cancer and drug resistance.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

We wish to thank Mr. Sambit Ghosh for reading the manuscript and for making useful comments. SV thanks the Council of Scientific and Industrial Research (CSIR), Government of India, for Emeritus Fellowship. We acknowledge Department of Biotechnology, Government of India for Fellowship to SG and for computing facility at MBU, IISc. Microsoft Research, India is also acknowledged for supporting the research.

REFERENCES

[1] (a) Vidal, M.; Cusick, M.E.; Barabasi, A.-L. Interactome networks and human disease. *Cell,* **2011,** *144*(6), 986-998; (b) Csermely, P.; Korcsmáros, T.; Kiss, H. J.; London, G.; Nussinov, R. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Ther.,* **2013,** *138*(3), 333-408; (c) Atilgan, C.; Okan, O.B.; Atilgan, A.R. Network-based models as tools hinting at nonevident protein functionality. *Ann. Rev. Biophys.,* **2012,** *41*, 205-225; (d) Raman, K. Construction and analysis of protein–protein interaction networks. *Autom. Exp.,* **2010**, *2*, 2.

[2] (a) Di Paola, L.; De Ruvo, M.; Paci, P.; Santoni, D.; Giuliani, A. Protein contact networks: an emerging paradigm in chemistry. *Chem. Rev.,* **2012,** *113*(3), 1598-1613; (b) Vishveshwara, S.; Ghosh, A.; Hansia, P. Intra and inter-molecular communications through protein structure network. *Curr. Protein Pept. Sci.,* **2009,** *10*(2), 146-160; (c) Vishveshwara, S.; Brinda, K.; Kannan, N. Protein structure: insights from graph theory. *J. Theor. Comput. Chem.,* **2002,** *1*(01), 187-211; (d) Yamada, T.; Bork, P. Evolution of biomolecular networks—lessons from metabolic and protein interactions. *Nat. Rev. Mol. Cell Biol.,* **2009,** *10*(11), 791-803.

[3] (a) Bahar, I.; Atilgan, A.R.; Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.,* **1997,** *2*(3), 173-181; (b) Atilgan, A.; Durell, S.; Jernigan, R.; Demirel, M.; Keskin, O.; Bahar, I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.,* **2001,** *80*(1), 505-515.

[4] Kannan, N.; Vishveshwara, S. Identification of side-chain clusters in protein structures by a graph spectral method. *J. Mol. Biol.,* **1999,** *292*(2), 441-464.

[5] Grindley, H.M.; Artymiuk, P.J.; Rice, D.W.; Willett, P. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol.,* **1993,** *229*(3), 707-721.

[6] Atilgan, A.R.; Atilgan, C. Local motifs in proteins combine to generate global functional moves. *Brief. Funct. Genomics,* **2012**, *11*(6), 479-488.

[7] (a) Bagler, G.; Sinha, S. Network properties of protein structures. *Physica A,* **2005,** *346*(1), 27-33; (b) Patra, S.; Vishveshwara, S. Backbone cluster identification in proteins by a graph theoretical method. *Biophys. Chem.,* **2000,** *84*(1), 13-25.

[8] (a) Bhattacharyya, M.; Vishveshwara, S. Probing the allosteric mechanism in pyrrolysyl-tRNA synthetase using energy-weighted network formalism. *Biochemistry,* **2011,** *50*(28), 6225-6236; (b) Vijayabaskar, M.S.; Vishveshwara, S. Interaction energy based protein structure networks. *Biophys. J.,* **2010,** *99*(11), 3704-15.

[9] Basu, S.; Bhattacharyya, D.; Banerjee, R. Mapping the distribution of packing topologies within protein interiors shows predominant preference for specific packing motifs. *BMC Bioinformatics,* **2011,** *12*(1), 195.

[10] Selvaratnam, R.; Chowdhury, S.; VanSchouwen, B.; Melacini, G. Mapping allostery through the covariance analysis of NMR chemical shifts. *Proc. Natl. Acad. Sci. U. S. A.,* **2011,** *108*(15), 6133-6138.

[11] (a) Bhattacharyya, M.; Upadhyay, R.; Vishveshwara, S. Interaction signatures stabilizing the nad (p)-binding rossmann fold: a structure network approach. *PLoS One,* **2012,** *7*(12), e51676; (b) Greene, L.H.; Hamada, D.; Eyles, S.J.; Brew, K. Conserved signature proposed for folding in the lipocalin superfamily. *FEBS Lett.,* **2003,** *553*(1), 39-44.

[12] Sethi, A.; Eargle, J.; Black, A.A.; Luthey-Schulten, Z. Dynamical networks in tRNA: protein complexes. *Proc. Natl. Acad. Sci. U. S. A.,* **2009,** *106*(16), 6620-6625.

[13] Miyazawa, S.; Jernigan, R.L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.,* **1996,** *256*(3), 623-44.

[14] Feynman, R. *The Feynman lectures on physics*. Addison-Wesley Publishing Company: Reading, MA, **1963**.

[15] Cooper, A.; Dryden, D. Allostery without conformational change. *Eur. Biophys. J.,* **1984,** *11*(2), 103-109.

[16] Greene, L.H. Protein structure networks. *Brief. Funct. Genomics,* **2012,** *11*(6), 469-478.

[17] (a) Sukhwal, A.; Bhattacharyya, M.; Vishveshwara, S. Network approach for capturing ligand-induced subtle global changes in protein structures. *Acta Crystallogr. D Biol. Crystallogr.,* **2011,** *67*(5), 429-439; (b) Brinda, K.; Vishveshwara, S. A network representation of protein structures: implications for protein stability. *Biophys. J.,* **2005,** *89*(6), 4159-4170.

[18] Sathyapriya, R.; Vijayabaskar, M.; Vishveshwara, S. Insights into protein–DNA interactions through structure network analysis. *PLoS Comput. Biol.,* **2008,** *4*(9), e1000170.

[19] Ghosh, S.; Chandra, N.; Vishveshwara, S. Investigating the Mechanism of Iron Dependent Repressor (IDER) Activation and DNA Binding. *Biophys. J.,* **2015,** *108*(2), 377a.

[20] Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A.E.; Berendsen, H.J. GROMACS: fast, flexible, and free. *J. Comput. Chem.,* **2005,** *26*(16), 1701-1718.

[21] Case, D.A.; Cheatham, T.E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K.M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R.J. The Amber biomolecular simulation programs. *J. Comput. Chem.,* **2005,** *26*(16), 1668-1688.

[22] Brooks, B.R.; Brooks, C.L.; MacKerell, A.D.; Nilsson, L.; Petrella, R.J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S. CHARMM: the biomolecular simulation program. *J. Comput. Chem.,* **2009,** *30*(10), 1545-1614.

[23] (a) Ghosh, A.; Vishveshwara, S. Variations in Clique and Community Patterns in Protein Structures during Allosteric Communication: Investigation of Dynamically Equilibrated Structures of Methionyl tRNA Synthetase Complexes†. *Biochemistry,* **2008,** *47*(44), 11398-11407; (b) Ghosh, A.; Vishveshwara, S. A study of communication pathways in methionyl-tRNA synthetase by molecular dynamics simulations and structure network analysis. *Proc. Natl. Acad. Sci. U. S. A.,* **2007,** *104*(40), 15711-15716.

[24] (a) Keskin, O.; Ma, B.; Nussinov, R. Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol.,* **2005,** *345*(5), 1281-1294; (b) del Sol, A.; O'Meara, P. Small‐world network approach to identify key residues in protein–protein interaction. *Proteins,* **2005,** *58*(3), 672-682.

[25] (a) Ma, B.; Elkayam, T.; Wolfson, H.; Nussinov, R. Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl. Acad. Sci. U. S. A.,* **2003,** *100*(10), 5772-5777; (b) Süel, G.M.; Lockless, S.W.; Wall, M.A.; Ranganathan, R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.,* **2003,** *10*(1), 59-69.

[26] (a) Brinda, K.; Surolia, A.; Vishveshwara, S. Insights into the quaternary association of proteins through structure graphs: a case study of lectins. *Biochem. J.,* **2005,** *391*, 1-15; (b) Sathyapriya, R.; Brinda, K.; Vishveshwara, S. Correlation of the side-chain hubs with the functional residues in DNA binding protein structures. *J. Chem. Inf. Model.,* **2006,** *46*(1), 123-129; (c) del Sol, A.; Fujihashi, H.; Amoros, D.; Nussinov, R. Residue centrality, functionally important residues, and active site shape: Analysis of enzyme and non‐enzyme families. *Protein Sci.,* **2006,** *15*(9), 2120-2128; (d) Böde, C.; Kovács, I.A.; Szalay, M.S.; Palotai, R.; Korcsmáros, T.; Csermely, P. Network analysis of protein dynamics. *FEBS Lett.,* **2007,** *581*(15), 2776-2782; (e) Amitai, G.; Shemesh, A.; Sitbon, E.; Shklar, M.; Netanely, D.; Venger, I.; Pietrokovski, S. Network analysis of protein structures identifies functional residues. *J. Mol. Biol.,* **2004,** *344*(4), 1135-1146.

[27] Brinda, K.; Vishveshwara, S. Oligomeric protein structure networks: insights into protein-protein interactions. *BMC Bioinformatics,* **2005,** *6*(1), 296.

[28] Watts, D.J.; Strogatz, S.H. Collective dynamics of 'small-world'networks. *Nature,* **1998,** *393*(6684), 440-442.

[29] (a) Maslov, S.; Sneppen, K. Specificity and stability in topology of protein networks. *Science,* **2002,** *296*(5569), 910-913; (b) Newman, M.E. Assortative mixing in networks. *Phys. Rev. Lett.,* **2002,** *89*(20), 208701.

[30] Bagler, G.; Sinha, S. Assortative mixing in protein contact networks and protein folding kinetics. *Bioinformatics,* **2007,** *23*(14), 1760-1767.

[31] Cormen, T.H.; Leiserson, C.E.; Rivest, R.L.; Stein, C. Depth-first search. *Introduction to algorithms,* **2001**, 540-549.

[32] (a) Stelzl, U.; Worm, U.; Lalowski, M.; Haenig, C.; Brembeck, F.H.; Goehler, H.; Stroedicke, M.; Zenkner, M.; Schoenherr, A.; Koeppen, S. A human protein-protein interaction network: a resource for annotating the proteome. *Cell,* **2005,** *122*(6), 957-968; (b) Jeong, H.; Mason, S.P.; Barabási, A.-L.; Oltvai, Z.N. Lethality and centrality in protein networks. *Nature,* **2001,** *411*(6833), 41-42.

[33] Sistla, R.K.; Brinda, K.; Vishveshwara, S. Identification of domains and domain interface residues in multidomain proteins from graph spectral method. *Proteins,* **2005,** *59*(3), 616-626.

[34] Erdős, P.; Rényi, A. On the strength of connectedness of a random graph. *Acta Math. Hung.,* **1961,** *12*(1), 261-267.

[35] Chatterjee, S.; Ghosh, S.; Vishveshwara, S. Network properties of decoys and CASP predicted models: a comparison with native protein structures. *Mol. BioSyst.,* **2013,** *9*(7), 1774-1788.

[36] Deb, D.; Vishveshwara, S.; Vishveshwara, S. Understanding protein structure from a percolation perspective. *Biophys. J.,* **2009,** *97*(6), 1787-1794.

[37] Adamcsek, B.; Palla, G.; Farkas, I.J.; Derényi, I.; Vicsek, T. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics,* **2006,** *22*(8), 1021-1023.

[38] Bhattacharyya, M.; Ghosh, A.; Hansia, P.; Vishveshwara, S. Allostery and conformational free energy changes in human tryptophanyl‐tRNA synthetase from essential dynamics and structure networks. *Proteins,* **2010,** *78*(3), 506-517.

[39] Chatterjee, S.; Bhattacharyya, M.; Vishveshwara, S. Network properties of protein-decoy structures. *J. Biomol. Struct. Dyn.,* **2012,** *29*(6), 1110-1126.

[40] Barabási, A.-L.; Albert, R. Emergence of scaling in random networks. *Science,* **1999,** *286*(5439), 509-512.

[41] Milenković, T.; Filippis, I.; Lappe, M.; Pržulj, N. Optimized null model for protein structure networks. *PLoS One,* **2009,** *4*(6), e5967.

[42] (a) Wolf, Y.I.; Grishin, N.V.; Koonin, E.V. Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.,* **2000,** *299*(4), 897-905; (b) Chothia, C. Proteins. One thousand families for the molecular biologist. *Nature,* **1992,** *357*(6379), 543.

[43] Noble, W.S. What is a support vector machine? *Nat. Biotechnol.,* **2006,** *24*(12), 1565-1567.

[44] Park, K.-J.; Kanehisa, M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics,* **2003,** *19*(13), 1656-1663.

[45] (a) Furey, T.S.; Cristianini, N.; Duffy, N.; Bednarski, D.W.; Schummer, M.; Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics,* **2000,** *16*(10), 906-914; (b) Khan, J.; Wei, J.S.; Ringner, M.; Saal, L.H.; Ladanyi, M.; Westermann, F.; Berthold, F.; Schwab, M.; Antonescu, C.R.; Peterson, C. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.,* **2001,** *7*(6), 673-679.

[46] Cai, Y.-D.; Liu, X.-J.; Xu, X.-b.; Chou, K.-C. Prediction of protein structural classes by support vector machines. *Comput. Chem.,* **2002,** *26*(3), 293-296.

[47] Bradford, J.R.; Westhead, D.R. Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics,* **2005,** *21*(8), 1487-1494.

[48] (a) Dong, Q.; Chen, Y.; Zhou, S. A machine learning-based method for protein global model quality assessment. *Int. J. Gen. Syst.,* **2011,** *40*(04), 417-425; (b) Kryshtafovych, A.; Fidelis, K. Protein structure prediction and model quality assessment. *Drug Discov. Today,* **2009,** *14*(7), 386-393.

[49] (a) Skiena, S. Dijkstra's Algorithm. *Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica, Reading, MA: Addison-Wesley* **1990**, 225-227; (b) Floyd, R.W. Algorithm 97: shortest path. *Commun. ACM.,* **1962,** *5*(6), 345.

[50] Ghosh, A.; Sakaguchi, R.; Liu, C.; Vishveshwara, S.; Hou, Y.-M. Allosteric Communication in Cysteinyl tRNA Synthetase a network of direct and indirect readout. *J. Biol. Chem.,* **2011,** *286*(43), 37721-37731.

[51] Daily, M.D.; Gray, J.J. Allosteric communication occurs via networks of tertiary and quaternary motions in proteins. *PLoS Comput. Biol.,* **2009,** *5*(2), e1000293.

[52] Blacklock, K.; Verkhivker, G.M. Computational modeling of allosteric regulation in the Hsp90 chaperones: a statistical ensemble analysis of protein structure networks and allosteric communications. *PLoS Comput. Biol.,* **2014,** *10*(6), e1003679.

[53] Amaro, R.E.; Sethi, A.; Myers, R.S.; Davisson, V.J.; Luthey-Schulten, Z.A. A network of conserved interactions regulates the allosteric signal in a glutamine amidotransferase. *Biochemistry,* **2007,** *46*(8), 2156-2173.

[54] (a) Tzeng, S.-R.; Kalodimos, C.G. Dynamic activation of an allosteric regulatory protein. *Nature,* **2009,** *462*(7271), 368-372; (b) Popovych, N.; Sun, S.; Ebright, R.H.; Kalodimos, C.G. Dynamically driven protein allostery. *Nat. Struct. Mol. Biol.,* **2006,** *13*(9), 831-838.

[55] del Sol, A.; Tsai, C.-J.; Ma, B.; Nussinov, R. The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure,* **2009,** *17*(8), 1042-1050.

[56] (a) Eyal, E.; Chennubhotla, C.; Yang, L.-W.; Bahar, I. Anisotropic fluctuations of amino acids in protein structures: insights from X-ray crystallography and elastic network models. *Bioinformatics,* **2007,** *23*(13), i175-i184; (b) Wang, Y.; Rader, A.; Bahar, I.; Jernigan, R.L. Global ribosome motions revealed with elastic network model. *J. Struct. Biol.,* **2004,** *147*(3), 302-314; (c) Bahar, I.; Rader, A. Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.,* **2005,** *15*(5), 586-592.

[57] Mitchell, E.M.; Artymiuk, P.J.; Rice, D.W.; Willett, P. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J. Mol. Biol.,* **1990,** *212*(1), 151-166.

[58] Halabi, N.; Rivoire, O.; Leibler, S.; Ranganathan, R. Protein sectors: evolutionary units of three-dimensional structure. *Cell,* **2009,** *138*(4), 774-786.

[59] (a) Vijayabaskar, M.; Vishveshwara, S. Insights into the fold organization of TIM barrel from interaction energy based structure networks. *PLoS Comput. Biol.,* **2012,** *8*(5), e1002505; (b) Soundararajan, V.; Raman, R.; Raguram, S.; Sasisekharan, V.; Sasisekharan, R. Atomic interaction networks in the core of protein domains and their native folds. *PLoS One,* **2010,** *5*(2), e9391.

[60] Brinda, K.; Vishveshwara, S.; Vishveshwara, S. Random network behaviour of protein structures. *Mol. Biosyst.,* **2010,** *6*(2), 391-398.

[61] De Ruvo, M.; Giuliani, A.; Paci, P.; Santoni, D.; Di Paola, L. Shedding light on protein–ligand binding by graph theory: The topological nature of allostery. *Biophys. Chem.,* **2012,** *165*, 21-29.

[62] Cui, Q.; Karplus, M. Allostery and cooperativity revisited. *Protein Sci.,* **2008,** *17*(8), 1295-1307.

[63] Monod, J.; Wyman, J.; Changeux, J.-P. On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.,* **1965,** *12*(1), 88-118.

[64] Koshland Jr, D.; Nemethy, G.; Filmer, D. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry,* **1966,** *5*(1), 365-385.

[65] (a) Perutz, M.F.; Wilkinson, A.; Paoli, M.; Dodson, G. The stereochemical mechanism of the cooperative effects in hemoglobin revisited. *Annu. Rev. Biophys. Biomol. Struct.,* **1998,** *27*(1), 1-34; (b) Perutz, M. Stereochemistry of cooperative effects in haemoglobin: haem–haem interaction and the problem of allostery. *Nature*, **1970,** *228*, 726-734; (c) Horovitz, A.; Willison, K.R. Allosteric regulation of chaperonins. *Curr. Opin. Struct. Biol.,* **2005,** *15*(6), 646-651.

[66] Frauenfelder, H.; McMahon, B.H.; Austin, R.H.; Chu, K.; Groves, J.T. The role of structure, energy landscape, dynamics, and allostery in the enzymatic function of myoglobin. *Proc. Natl. Acad. Sci. U. S. A.,* **2001,** *98*(5), 2370-2374.

[67] (a) Conn, P.J.; Christopoulos, A.; Lindsley, C.W. Allosteric modulators of GPCRs: a novel approach for the treatment of CNS disorders. *Nat. Rev. Drug Discov.,* **2009,** *8*(1), 41-54; (b) Christopoulos, A. Allosteric binding sites on cell-surface receptors: novel targets for drug discovery. *Nat. Rev. Drug Discov.,* **2002,** *1*(3), 198-210.

[68] Marvin, J.; Corcoran, E.; Hattangadi, N.; Zhang, J.; Gere, S.; Hellinga, H. The rational design of allosteric interactions in a monomeric protein and its applications to the construction of biosensors. *Proc. Natl. Acad. Sci. U. S. A.,* **1997,** *94*(9), 4366-4371.

[69] Motlagh, H.N.; Wrabl, J.O.; Li, J.; Hilser, V.J. The ensemble nature of allostery. *Nature,* **2014,** *508*(7496), 331-339.

[70] (a) Motlagh, H.N.; Hilser, V.J. Agonism/antagonism switching in allosteric ensembles. *Proc. Natl. Acad. Sci. U. S. A.,* **2012,** *109*(11), 4134-4139; (b) Hilser, V.J. An ensemble view of allostery. *Science,* **2010,** *327*(5966), 653-654; (c) Boehr, D.D.; Nussinov, R.; Wright, P.E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.,* **2009,** *5*(11), 789-796; (d) Gunasekaran, K.; Ma, B.; Nussinov, R. Is allostery an intrinsic property of all dynamic proteins? *Proteins,* **2004,** *57*(3), 433-443;(e) Tsai, C.-J.; Nussinov, R. A unified view of "How Allostery Works". *PLoS Comput. Biol.,* **2014,** *10*(2), e1003394.

[71] Goodey, N.M.; Benkovic, S.J. Allosteric regulation and catalysis emerge via a common route. *Nat. Chem. Biol.,* **2008,** *4*(8), 474-482.

[72] (a) Özkan, E.; Yu, H.; Deisenhofer, J. Mechanistic insight into the allosteric activation of a ubiquitin-conjugating enzyme by RING-type ubiquitin ligases. *Proc. Natl. Acad. Sci. U. S. A.,* **2005,** *102*(52), 18890-18895; (b) Bahar, I.; Chennubhotla, C.; Tobi, D. Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. *Curr. Opin. Struct. Biol.,* **2007,** *17*(6), 633-640; (c) Palazzesi, F.; Barducci, A.; Tollinger, M.; Parrinello, M. The allosteric communication pathways in KIX domain of CBP. *Proc. Natl. Acad. Sci. U. S. A.,* **2013,** *110*(35), 14237-14242.

[73] Amadei, A.; Linssen, A.; Berendsen, H.J. Essential dynamics of proteins. *Proteins,* **1993,** *17*(4), 412-425.

[74] (a) Daidone, I.; Amadei, A.; Di Nola, A. Thermodynamic and kinetic characterization of a β−hairpin peptide in solution: An extended phase space sampling by molecular dynamics simulations in explicit water. *Proteins,* **2005,** *59*(3), 510-518; (b) Patel, S.; Sasidhar, Y.U. A shorter peptide model from staphylococcal nuclease for the folding–unfolding equilibrium of a β-hairpin shows that unfolded state has significant contribution from compact conformational states. *J. Struct. Biol.,* **2008,** *164*(1), 60-74.

[75] Nozawa, K.; O'Donoghue, P.; Gundllapalli, S.; Araiso, Y.; Ishitani, R.; Umehara, T.; Söll, D.; Nureki, O. Pyrrolysyl-tRNA synthetase–tRNAPyl structure reveals the molecular basis of orthogonality. *Nature,* **2009,** *457*(7233), 1163-1167.

[76] Fersht, A.R. Demonstration of two active sites on a monomeric aminoacyl-tRNA synthetase. Possible roles of negative cooperativity and half-of-the-sites reactivity in oligomeric enzymes. *Biochemistry,* **1975,** *14*(1), 5-12.

[77] Pieniazek, S.N.; Hingorani, M.M.; Beveridge, D. Dynamical allosterism in the mechanism of action of DNA mismatch repair protein muts. *Biophys. J.,* **2011,** *101*(7), 1730-1739.

[78] (a) Anfinsen, C.B.; Haber, E. Studies on the reduction and reformation of protein disulfide bonds. *J. Biol. Chem.,* **1961,** *236*(5), 1361-1363; (b) Anfinsen, C.B. Principles that govern the folding of protein chains. *Science,* **1973,** *181*(4096), 223-230.

[79] Kryshtafovych, A.; Moult, J.; Bales, P.; Bazan, J.F.; Biasini, M.; Burgin, A.; Chen, C.; Cochran, F.V.; Craig, T.K.; Das, R. Challenging the state of the art in protein structure prediction: Highlights of experimental target structures for the 10th Critical Assessment of Techniques for Protein Structure Prediction Experiment CASP10. *Proteins,* **2014,** *82*(S2), 26-42.

[80] Ghosh, S.; Vishveshwara, S. Ranking the quality of protein structure models using sidechain based network properties. *F1000Research* **2014,** *3*, 17.

[81] (a) Ghosh, S.; Ghosh, S.; Vishveshwara, S.; WeFold-Community In *Quality assessment of protein structure models using Protein Structure Networks [PSN-QA]*, CASP11 Meeting, Mexico, Mexico, **2015**; (b) Khatib, F.; Shah, A.; Trieber, N.; Koepnick, B.; Flatten, J.; Cooper, S.; Chapman, N.; Popović, Z.; Baker, D.; Players, F.; Wallner, B.; Chopra, G.; Levitt, M.; Ghosh, S.; Vishveshwara, S.; Crivelli, S. In *Selection of Foldit refinement models using PSN-QA*, CASP11 Meeting, Mexico, Mexico, **2015**.

[82] Van Wart, A.T.; Durrant, J.; Votapka, L.; Amaro, R.E. Weighted implementation of suboptimal paths (WISP): an optimized algorithm and tool for dynamical network analysis. *J. Chem. Theory Comput.,* **2014,** *10*(2), 511-517.

[83] Pasi, M.; Tiberti, M.; Arrigoni, A.; Papaleo, E. xPyder: a PyMOL plugin to analyze coupled residues and their networks in protein structures. *J. Chem. Inf. Model.,* **2012,** *52*(7), 1865-1874.

[84] Eargle, J.; Luthey-Schulten, Z. NetworkView: 3D display and analysis of protein· RNA interaction networks. *Bioinformatics,* **2012,** *28*(22), 3000-3001.

[85] Vijayabaskar, M.; Niranjan, V.; Vishveshwara, S. GraProStr—graphs of protein structures: a tool for constructing the graphs and generating graph parameters for protein structures. *Open Bioinform. J.,* **2011,** *5*(2011), 6.

[86] Bhattacharyya, M.; Bhat, C.R.; Vishveshwara, S. An automated approach to network features of protein structure ensembles. *Protein Sci.,* **2013,** *22*(10), 1399-1416.

[87] (a) Krissinel, E.; Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.,* **2004,** *60*(12), 2256-2268; (b) Koch, I.; Kaden, F.; Selbig, J. Analysis of protein sheet topologies by graph theoretical methods. *Proteins,* **1992,** *12*(4), 314-323.

[88] Jiao, X.; Chang, S.; Li, C.-h.; Wang, C.-x. Construction and application of the weighted amino acid network based on energy. *Phys. Rev. E,* **2007,** *75*(5), 051903.

[89] Artymiuk, P.J.; Poirrette, A.R.; Grindley, H.M.; Rice, D.W.; Willett, P. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J. Mol. Biol.,* **1994,** *243*(2), 327-344.

[90] (a) Kiefer, L.L.; Paterno, S.A.; Fierke, C.A. Hydrogen bond network in the metal binding site of carbonic anhydrase enhances zinc affinity and catalytic efficiency. *J. Am. Chem. Soc.,* **1995,** *117*(26), 6831-6837; (b) Taraphder, S.; Hummer, G. Protein side-chain motion and hydration in proton-transfer pathways. Results for cytochrome P450cam. *J. Am. Chem. Soc.,* **2003,** *125*(13), 3931-3940.

[91] Soffer, S.N.; Vazquez, A. Network clustering coefficient without degree-correlation biases. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.,* **2005,** *71*(5), 057101.

Received: March 18, 2015 **Revised: May 17, 2015** **Accepted: May 21, 2015**