Computational and Structural Biotechnology Journal 14 (2016) 245–251

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Biophysical and computational methods to analyze amino acid interaction networks in proteins

Kathleen F. O'Rourke [1], Scott D. Gorman [1], David D. Boehr *

Department of Chemistry, The Pennsylvania State University, University Park, PA 16802, USA

ARTICLE INFO

Article history:
Received 30 March 2016
Received in revised form 4 June 2016
Accepted 13 June 2016
Available online 22 June 2016

Keywords:
Amino acid interaction network
Allostery
Graph theory
CONTACT
Molecular dynamics
Elastic network
NMR
Coevolution
Statistical coupling analysis

ABSTRACT

Globular proteins are held together by interacting networks of amino acid residues. A number of different structural and computational methods have been developed to interrogate these amino acid networks. In this review, we describe some of these methods, including analyses of X-ray crystallographic data and structures, computer simulations, NMR data, and covariation among protein sequences, and indicate the critical insights that such methods provide into protein function. This information can be leveraged towards the design of new allosteric drugs, and the engineering of new protein function and protein regulation strategies.

1. Introduction

It has long been understood that interactions at the local level (e.g. H-bonding, steric interactions) dictate the formation of protein structural elements, such as α-helices and β-sheets, and that local interactions also dictate the packing of these various structural elements to form three-dimensional protein structure (e.g. ref. [1,2]). There is also now a better appreciation for the local interactions that are important for loop structure and dynamics (e.g. ref. [3]). With these energetic considerations in mind, globular proteins can be viewed as being held together by a series of local interactions through networks of interacting amino acid residues. These amino acid networks (Fig. 1) have also been termed 'residue interaction networks' [4], 'protein structure networks' [5], 'contact networks' [6], 'pathways' [7], 'circuits' [8], 'wiring diagrams' [9], 'protein sectors' [10] and so on. Intrinsic to this viewpoint is the idea that some interactions and amino acid residues are more important than others, such that the amino acid network generally represents a subset of all potential interactions and residues within a protein. In some cases, there may be multiple amino acid networks identified (e.g. ref. [11]), where local changes primarily affect the interactions between the amino acid residues involved in a particular network.

A variety of diverse structural and computational methods have been developed to delineate amino acid networks in proteins, and these methods have provided tremendous insights into protein function. In this Review, we highlight some of the different computational and experimental methods that have been used to delineate amino acid networks in proteins (Table 1), and indicate the insight that these approaches have given regarding protein function. We note that other recent review articles have been written on many of these methods, including graph theory [6], molecular dynamics (MD) simulations [12], elastic network models (ENM [13]), NMR methods to study allostery and amino acid networks [14] and bioinformatics methods to identify co-evolving residues [15], and as such, we do not treat these methods comprehensively. We also recognize that the length of this review prevents us from being exhaustive with our examples.

2. Network approaches to understanding protein function

In biology, network interactions have been analyzed from the species to the molecular level [16–18]. The elegance of this mathematical theory is to simplify a complex problem into a set of nodes and edges,

* Corresponding author at: 107 Chemistry Building, The Pennsylvania State University, University Park, PA, USA.
E-mail address: ddb12@psu.edu (D.D. Boehr).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.csbj.2016.06.002

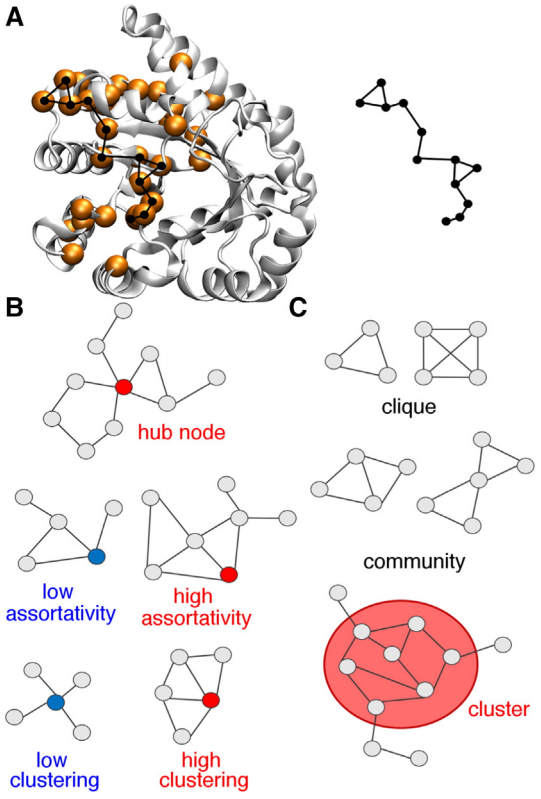

Fig. 1. Proteins can be viewed as interacting networks of amino acid residues. A. Partial network in the alpha subunit of tryptophan synthase (PDB 1K3U) identified by NMR methods [93]. In the network representation, the nodes are the amino acid residues and represented by circles, and the edges are interactions between the residues and are indicated by lines joining the circles. B. Concepts related to network theory, including hub residues, assortativity and clustering. C. Networks can follow a hierarchy of connectivities, ranging from smaller cliques to larger clusters. Panels B and C were adapted from ref. [38].

together known as a 'graph' [19–21]. Graphical approaches have provided intuitive pictures and useful insights for analyzing many complex biological problems, including enzyme-catalyzed reactions [22–24], inhibition of HIV-1 reverse transcriptase [25], inhibition kinetics of processive nucleic acid polymerases and nucleases [26], protein folding kinetics [27] and drug metabolism systems [28]. In the context of protein structure, the amino acid side-chains, or whole amino acid residues, are most commonly treated as the nodes. An edge represents some type of interaction between two nodes. Edges can have a range of definitions such as the calculated energy of interaction, evolutionary conservation, or surface overlap [29–32]. An important feature of edges is the weighting, which may allocate different strengths to different types of interactions and/or provide a particular cut-off distance for residues in close sequence space [33]. There are many algorithms available to construct and analyze amino acid networks using graph theory, including CSU software [34], xPyder [35], PSN-Ensemble [36] and NetworkAnalyzer [37].

Other parameters of the protein graph may be used to further analyze the network, and be related to different structural and functional properties of the protein. For example, a 'hub node' has a higher number of edges connected to it than other nodes [38] (Fig. 1). Residues corresponding to hub nodes may be key factors for maintaining structure and determining function. For example, a large experimental set of T4 lysozyme protein variants was studied, where some amino acid substitutions had little to no effect on the function of the enzyme and some substitutions inactivated the protein [39]. All of the deleterious substitutions were later identified as central hub nodes [40].

Connectivity is an important feature of a protein graph. The clustering coefficient, C_v, provides a measure of connectivity through Eq. (1):

$$C_v = \frac{2e_v}{k_v(k_v-1)} \tag{1}$$

where k_v is the number of neighbors to node v, and e_v is the number of connected pairs among v neighbors. Residues that have a high connectivity are typically linked to separate clusters or communities of residues [38]. The assortativity matrix is another parameter that helps determine the impact a node has on the network (Fig. 1). This matrix is a measure of the number of connections between nodes. A more 'resilient' network [41] would have a higher assortativity, providing multiple paths to connect distant regions of the protein.

A group of nodes can be classified into different types according to how a signal might be transmitted through them (Fig. 1). A clique (or k-clique) is a complete subgraph, meaning that it is a set of nodes and edges that are connected to every other node in the subgraph [38]. Similar to cliques are communities, which represent a set of connected cliques [38]. Inspection of the cliques and communities in a given protein might be used to track small ligand-induced conformational changes and signal transmission, which can be indicative of the interaction strength of the effector molecule and the quality of the network as a whole. For example, differences in the cliques and communities between the apo and ligand bound states of methionyl t-RNA synthetase were used to understand inter-domain signaling [42]. The binding of ATP induces the formation of new cliques that allow for communication between distal areas of the enzyme.

A cluster has more relaxed requirements than a clique or a community (Fig. 1). In a cluster, the nodes have a higher connectivity with each other than with nodes outside the group, but not all interact pairwise [38]. The largest cluster may be important in defining the core of the protein and can involve up to 80% of the nodes in the entire network. For example, identification of hydrophobic subclusters was used to understand long-range interactions important for stabilizing the tertiary fold of proteins [43]. In this study, it was found that the clusters were larger in thermophilic proteins, which may lead to higher temperature stability.

Measures of residue centrality, including closeness (C_n) and betweenness, are often used to predict residues important for the transmission of information across a protein structure. The closeness centrality [44] is defined according to Eq. (2):

$$C_n = \frac{j-1}{\sum_{i\neq n} sd(i,n)} \tag{2}$$

where $sd(i,n)$ is the shortest path between nodes i and n, and j is the number of nodes in the network. The betweenness centrality B is determined as the fraction of shortest paths that pass through a node [45]. Residues with high C_n or B have been shown to play critical roles in protein function [40,46,47]. Other measures of residue centrality have been proposed (e.g. ref. [48,49] and references therein). Recent examples using these types of approaches include studies analyzing allosteric pathways in tRNA synthetases [50], G-protein coupled receptors [51], Hsp90 [52] and cyclophilin A [53].

3. More sophisticated structure-based approaches to network analysis

Conformational fluctuations in proteins are important in mediating their biological functions. For example, *E. coli* dihydrofolate reductase (DHFR) must pass through multiple conformations as it proceeds through its catalytic cycle [54]. Smaller fluctuations, such as those in side chains, may be evident in X-ray diffraction data [55], though they may be ignored during the refinement process when producing a structural model. The qFit algorithm was developed to fit these alternative