GraProStr – Graphs of Protein Structures: A Tool for Constructing the Graphs and Generating Graph Parameters for Protein Structures

M.S. Vijayabaskar[#], Vidya Niranjan[#] and Saraswathi Vishveshwara*

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

Abstract: Protein structures can be represented as graphs/networks by defining the amino-acids as nodes and the non-covalent interactions as connections (edges). An analysis of such a graph provides valuable insights into the global structural properties, function, folding, and stability of proteins. Here we have created a webtool *GraProStr* to generate protein structure networks and analyze network parameters. Protein side-chain based, Cα/Cβ backbone based or protein-ligand Graphs/Networks can be generated using *GraProStr*. The well tested tool is now made available to the scientific community for the first time. *GraProStr* is available online and can be accessed from http://vishgraph.mbu.iisc.ernet.in/GraProStr/index.html using any of the internet browsers (best viewed in Mozilla Firefox version ≥ 3.6). The webtool is written using Perl CGI and available using Apache Webserver. With its customizable definitions of protein structure networks and well defined network parameters, *GraProStr* can be a very useful tool for both theoretical and experimental elucidation of protein structures.

Keywords: Protein Structure Network, Non-covalent Interactions, Clusters, Hubs, Cliques, Communities.

1. INTRODUCTION

Proteins adopt unique three dimensional structures to perform various cellular functions. Hence valuable information on the structure formation, stability and functional correlations can be obtained from the analyses of protein structures. A wealth of knowledge has been obtained by investigating the non-covalent, pair-wise interactions in protein structures. However, the interactions take place in the global context of the protein structure. Therefore an investigation involving pair-wise interactions alone is not sufficient to capture all the determinants of protein structures. This consideration has led to the development of Protein Structure Networks (PSNs).

In most PSNs, amino-acid residues are considered as nodes and non-covalent interactions between them are represented as edges (links), on the basis of distances between Cα-Cα or Cβ-Cβ, or side-chain atoms (or protein-ligand atoms) [1-4]. The extent of side-chain interactions can be further quantified (called Interaction strength) based on the number of atom-atom contacts, which takes into account the mutual orientation of the interacting residues [5]. Different PSNs can be constructed for the same structure for varying values of interaction strength, to describe different aspects of the structure. Diverse problems such as the identification of clusters of interacting residues at the active-site, residues important for folding and stability, stabilization of protein-protein interfaces and the level of global connectivity have been addressed by the analyses of PSNs [6-11]. Further, PSNs are also used in elucidating important problems such as allosteric communications within proteins [10, 12]. The construction of PSN, analysis of parameters and their implications in protein structure and function and stability are reviewed in detail in ref [5, 6].

Even though the analyses of structures as networks have introduced new areas and perspectives for studying proteins, a comprehensive tool for such exploration is not yet available for the scientific community. Here we introduce the webtool -*GraProStr* that is developed for the first time to analyze protein structures as networks. This tool allows us to represent protein structures as coarse-grain PSNs based on Cα/Cβ atomic distances or finer level PSNs by considering the interaction strength between the constituent amino-acids as discussed in the references cited above. The generated PSNs can then be used to identify clusters of interacting amino-acids (stabilizing regions), highly interacting amino-acids in the form of hubs in PSNs and rigid regions in the protein structures in terms of cliques and communities. We have introduced a method by which ligands can also be considered while constructing networks and hence giving us a useful method to study protein-ligand interactions. We believe that *GraProStr*, with its array of customizable PSNs and network parameters can be an extremely valuable to structural biologists in identifying structurally and functionally important regions in proteins, thus enabling further experimental characterization.

2. *GraProStr*

GraProStr is a webtool to analyze protein structures as networks, constructed on various considerations. The different networks (classified based on edge definition) that can be obtained are: Protein Sidechain (PScNs), Cα-Cα or Cβ-Cβ distance based (PcNs) and Protein-Ligand (PLNs) Networks. PScNs use a parameter called Interaction Strength (I_{ij}, Eq 1), PcNs use distance based parameter (d_{ij}, Eq 2) for defining edges. The PLN uses I_{ij} for intra-protein edges and uses d_{ij} (d_{max}=4.5Å) for protein-ligand edges. The adjacency matri-

*Address correspondence to this author at the Molecular Biophysics Unit, Indian Institute of Science, Banglore 560012, India; Tel: +91-80-22932611; Fax: +91-80-23600535; E-mail: sv@mbu.iisc.ernet.in

[#]Equal contribution.

1875-0362/11

ces are generated as given in Eq 3 for PScN and Eq 4 for PcN.

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

$$d_{ij} = \left| r_i^{C\alpha} - r_j^{C\alpha} \right| \text{ or } d_{ij} = \left| r_i^{C\beta} - r_j^{C\beta} \right| \quad (2)$$

$$A_{ij} = \begin{cases} 1, I_{ij} \geq I_{min} \\ 0, \text{ otherwise} \end{cases} \quad (3)$$

$$A_{ij} = \begin{cases} 1, d_{ij} \leq d_{max} \\ 0, \text{ otherwise} \end{cases} \quad (4)$$

$$D_i = \sum_{j=1}^{X} A_{ij} \quad (5)$$

where n_{ij} is the total number interacting atomic pairs between residues '*i*' and '*j*', N_i and N_j are their normalization values, I_{min} is the Interaction Strength cutoff [2], $r_i^{C\alpha}$ is the *3D* position of the Cα for residue '*i*', d_{max} is the distance cutoff and *A* is the adjacency matrix. If A_{ij} is 1 then an edge exists between '*i*' and '*j*'. D_i is the degree of the node '*i*', which is the total number of edge incident on that node and *X* is the total number of nodes/residues in the PSN (Eq 5).

The adjacency matrices thus generated are used as input in this webtool to generate network parameters such as amino-acid clusters (set of autonomously connected nodes), hubs (highly connected nodes), cliques (rigid regions in proteins) and communities (group of cliques). Clusters are evaluated using Depth First Search algorithm [6], hubs are evaluated using the degree (edges incident on each node) of the nodes, cliques and communities are evaluated using CFinder (http://cfinder.org/) [13]. Further, we provide the network in the form of adjacency matrix and also render them using GraphViz (http://www.graphviz.org/) for visualization. The input file (.sif format) for the graph visualization program Cytoscape (www.cytoscape.org) is also provided. In multimeric proteins, these parameters are evaluated at both the monomer and at the interface levels. Use of ligands (small molecules and metal ions) in PSNs (PLNs) is a novel step towards analyzing protein-ligand complexes. Water molecules and other small molecules like metal ions are known to be biologically relevant mostly mediating the functional activity of proteins. If, according to the PDB format, these molecules are annotated or represented as hetero-atoms ("HETATM", see PDB format and also Supplementary Information), they can be regarded in the construction of PLNs to study their structural and functional significance. However, it should be noted that the presence of a large number of water molecules in the PDB file may give rise to water clusters, in which case, the coordinates of functionally relevant water molecules should be chosen judiciously.

3. *GraProStr* CAPABILITIES

The web tool can be accessed through http://vishgraph.mbu.iisc.ernet.in/GraProStr/index.html (Fig. **1**). The parameter outputs can be downloaded as flat files or archives. The flat files can be viewed using text editors and the archive contents can be accessed using winrar (Windows), tar and gunzip utilites (Linux, MacOS). CFinder can be used as graphical interface for viewing cliques or communities. The input file for Cfinder is provided by our tool. The graphical representation of the network can be downloaded as png image.

3.a. Input Parameters

The left panel in Fig. (**1**) gives the different types of networks that can be generated using this web-suite. The X-ray structure coordinates of the protein (or protein-ligand) atoms can be uploaded in PDB [4] format or the four letter PDB code. The network parameters to be analyzed can be selected from the drop down menu (Fig. **2**). If the submitted protein is a multimer, the webtool can be used effectively to separate monomeric parameters and interface parameters.

Unless specified, default cutoff values are taken for the construction of different networks. The default distance cut off of 6.5 Å, corresponding to the radial distribution function in proteins (between Cα or Cβ atoms) is used for PcN. However, the users can choose higher values (6.5 Å to 10 Å) if required, keeping in mind the increase in the number of connections. In PScN a default value of I_{min} = 4% is given. At a low I_{min} (0%, representing weak interaction due to any single atom-atom contact between two residues) a large number of residues in the protein get connected yielding a single large

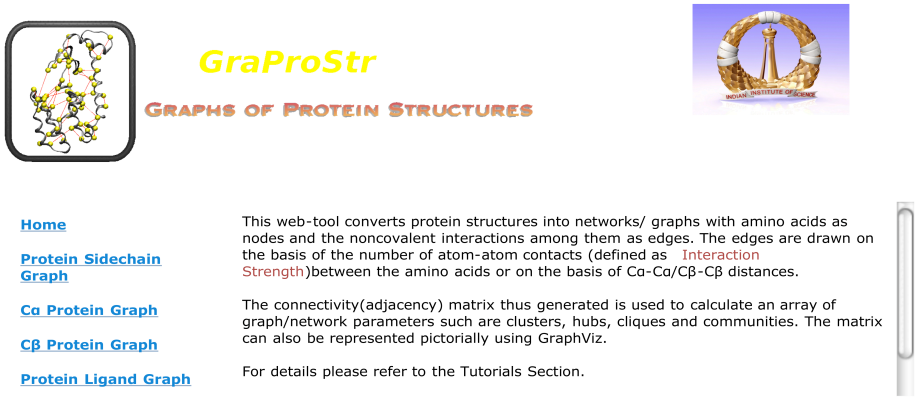

Fig. (1). *GraProStr* home page.

The above figure shows the *GraProStr* home page in which the different types of structure networks (PScN, PcN and PLN) that can be generated are provided as separate links in the menu (left).

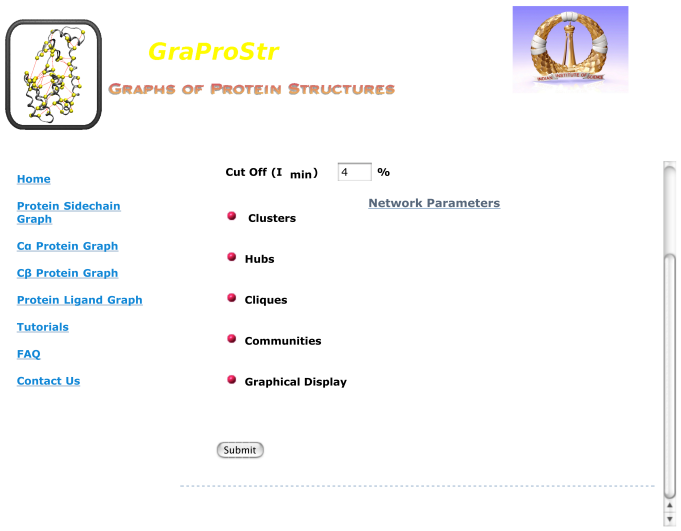

Fig. (2). Input Parameters for *GraProStr*.

The above figure highlights the various input parameters required by the webtool. The coordinate file (or PDB id) and the cutoff are essential parameters. The cutoff can be any real positive number. The network parameters like clusters, hubs, cliques, communities and graphical display are optional.

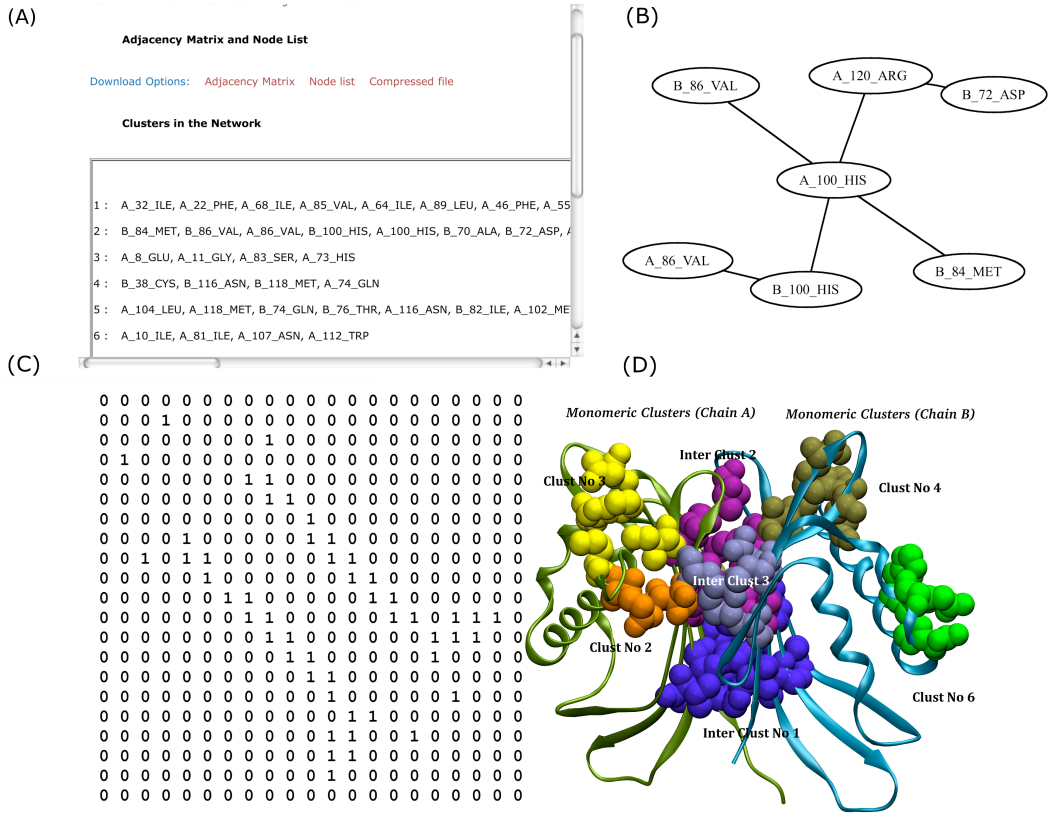

Fig. (3). Example output for the protein Nuclear Transport Protein (NTF2; PDB id – 1ask) and its clusters.

(**A**) The web snapshot of the output (adjacency matrix and graph) of a sample run using protein 1ask at 4% cutoff (PScN) is given. The cluster output is such that each cluster is given in a separate line and the nodes in a cluster are separated by commas. (**B**) A sub-graph of the structure network generated using GraphViz, where the ellipsoids represent the amino acid nodes and the solid lines represent the edges between them. (**C**) shows a section of the adjacency matrix. The rows and columns are the amino acids in the protein. If A_{ij}=0 then no edge exists between amino acids 'i' and 'j'. An edge exists if A_{ij}=1. (**D**) The clusters (both monomeric and interface) are represented as vdW representations. Each cluster is represented in a different color. There are six monomeric clusters (three from each chain, shown in (**A**)) and three interface clusters in the PScN of NTF2 at I_{min}=4%. It should be noted that we have not highlighted clusters 1 and 5 due to their large sizes and they encompass most of the residues in chains **A** and **B** respectively. The protein and the clusters have been rendered using Visual Molecular Dynamics (VMD) [17].

cluster. The transition from this stage to a number of smaller clusters takes place at I_{min} range of 2-4%. At I_{min}s greater than 4% (5-10%) yield highly interacting small clusters that may be involved in structure stabilization. The choice of I_{min} depends on the problem to be investigated, for instance, higher I_{min}s can be used for the identification of stabilizing clusters whereas lower I_{min}s can be used to identify cliques and communities, which are not found at higher I_{min} values.

3.b. Output Parameters

The tool generates the adjacency matrix (Fig. **3C**) and the corresponding node list along with the selected network parameters for the corresponding cutoff. The parameters are displayed in the form of tables and can also be downloaded as flat files or archives. The network generated using GraphViz is shown in Fig. (**3B**).

The network parameters are explained with a sample protein (Nuclear Transport Protein (NTF2); PDB id – 1ask).

Clusters

As mentioned above, in globular proteins, a single large cluster dominates a highly connected PSN (high d_{max} or low I_{min}). This cluster disintegrates into smaller, highly interact-

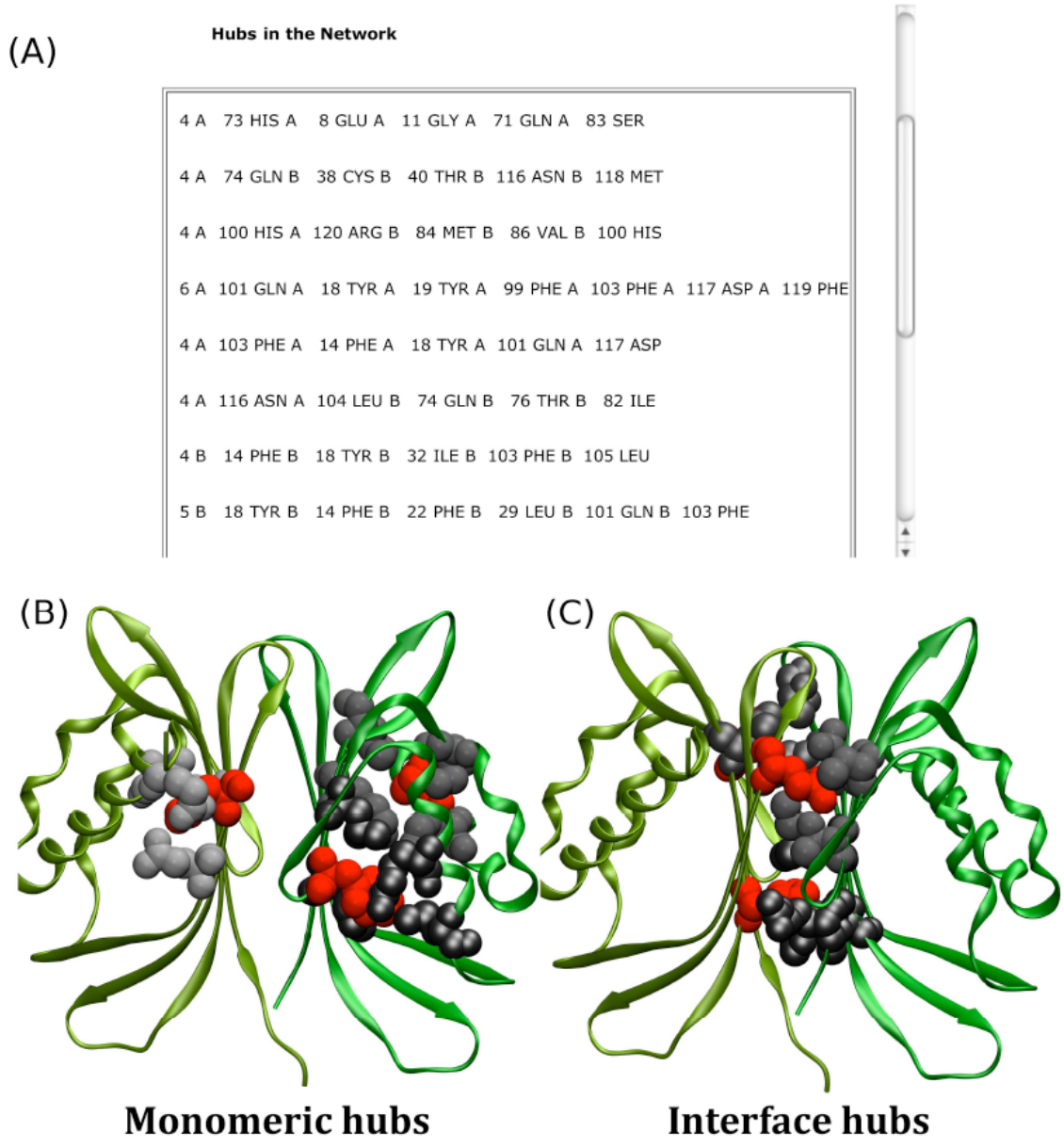

Fig. (4). Hubs in PScN of NTF2 and their neighbors.

(**A**) The above figure shows the hubs in 1ask (4% cutoff for PScN). The first column gives the total number of connections in the node (degree). It is followed by the hub residue (<chain> <resid> <resname>) and then its neighbors. The hubs along with its neighbours are highlighted as vdW spheres in (**B**) and (**C**). There are eight monomeric hubs (see (**A**)) and three interface hubs. Out of eight monomeric hubs three (H73, W41, F14) are highlighted (red) and their connections are given in different shades of gray in (**B**). The three interface hubs (Q74, H100 and N116) are also given (red) along with their neighbours (shades of gray) from the other interacting chain, in figure (**C**). Visual Molecular Dynamics (VMD) [17] was used for generating the figures of protein and its hubs.

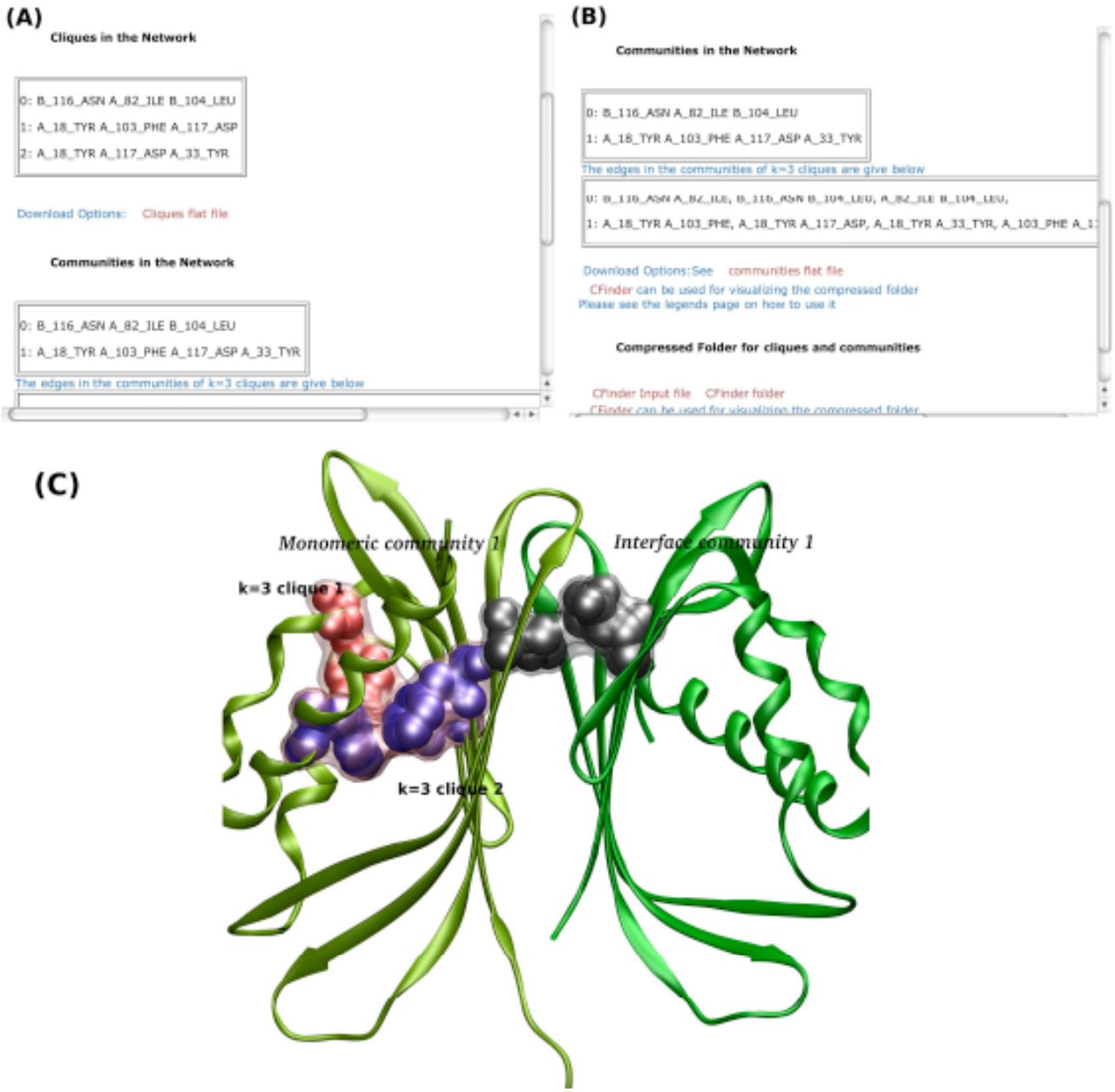

Fig. (5). Cliques and Communities in PScN of NTF2.

The above figure shows the cliques and communities in 1ask (PScN at 2% cutoff). (**A**) shows the format in which cliques and communities are displayed. Each line represents a unique clique/community and their members are separated by spaces. (**B**) Not all the edges are connected to each other in a community of cliques and hence the edge lists for the communities are also provided, where the edges are discriminated by commas and the nodes between which the edges are constructed are separated by spaces. (**C**) The cliques in the PScN of NTF2 (3%) are given. There are two monomeric *k=3* cliques (red vdW and blue vdW spheres) forming a community (pink transparent surf representation) with one clique (gray vdW spheres) forming an interface community (gray transparent surf representation). The protein and the cliques have been rendered using Visual Molecular Dynamics (VMD) [17].

ing clusters of amino-acids in a sparse PSN (low d_{max} or high I_{min}). Fig. (**3**) shows six clusters of different sizes, in PScN of NTF2 at cutoff=4%.

Hubs

Hubs are nodes whose degrees (D_i; Eq 5) are atleast 4. It has been observed that the population of hubs is high for a dense PSN, whereas only highly interacting hubs are present in sparse PSNs [14]. In Fig. (**4**), a list of hubs and their connections are given.

k=*n* *Clique* is a group of *'n'* nodes/residues with each node connected to every other node. A ***Community*** of k=*n* cliques is a collection of cliques in which *n-1* nodes are shared between two cliques (Fig. **5A**, **C**). The edges between nodes in a community are also given separately (Fig. **5B**). Cliques and communities are mostly present in highly connected PSNs.

A brief tutorial is given along with the web-tool and also as a supplementary material.

4. APPLICATIONS

The above mentioned network parameters have been used extensively to study the global structural properties, stability, and functions of proteins. For instance additional aromatic clusters which are implicated in the stability of thermophilic proteins [11], and interface clusters crucial for oligomerization have been identified [9, 15]. Hub residues important for protein structure stability and the functionally important ones around the active site have been detected [8, 9]. Cliques and communities which have been implicated for the robustness in long-range allosteric communication in systems like tRNA synthetases have also been identified [12, 16]. With a variety of structure networks and a set of easily interpretable network parameters, *GraProStr* can be a useful tool in structural biology.

Citation

The following reference should be cited by the users of this program:

GraProStr: Indian Institute of Science, Bangalore, India (2010) Saraswathi Vishveshwara, N. Kannan, Swarna Mayee Patra, Rakesh Kumar Pandey, K.V. Brinda, R. Sathyapriya, Amit Ghosh, M.S. Vijayabaskar and Vidya Niranjan.

ACKNOWLEDGEMENTS

The authors thank the Bioinformatics Center at the Indian Institute of Science, Bangalore, India for providing us with an easily accessible inhouse PDB repository. We thank the CFinder group for permitting us to use their software to generate cliques and communities. Supercomputer Education and Research Center (SERC), Indian Institute of Science, Bangalore, India is thanked for providing us with computational facilities. Department of Science and Technology, India (DST Mathematical Biology Grant, DST0773) is acknowledged for funding the project.

ABBREVIATIONS

PcN = Protein Cα/β Networks

PDB = Protein Data Bank

PLN = Protein Ligand Networks

PScN = Protein Side-chain Networks

PSN = Protein Structure Networks

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publishers Web site along with the published article.

REFERENCES

[1] G. Bagler, and S. Sinha, "Network properties of protein structures," *Physica A.*, vol. 346, pp. 27-33, 2004.

[2] L. H. Greene, and V. A. Higman, "Uncovering network systems within protein structures," *J. Mol. Biol.*, vol. 334, pp. 781-91, 2003.

[3] N. Kannan, and S. Vishveshwara, "Identification of side-chain clusters in protein structures by a graph spectral method," *J. Mol. Biol.*, vol. 292, pp. 441-64, 1999.

[4] M. Vendruscolo, N. V. Dokholyan, E. Paci, and M. Karplus, "Small-world view of the amino acids that play a key role in protein folding," *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.*, vol. 65, p. 061910, 2002.

[5] S. Vishveshwara, A. Ghosh, and P. Hansia, "Intra and intermolecular communications through protein structure network," *Curr. Protein. Pept. Sci.*, vol. 10, pp. 146-60, 2009.

[6] S. Vishveshwara, K. V. Brinda, and N. Kannan, " Protein Structure: Insights from Graph Theory," *J. Theo. Comp. Chem.*, vol. 1, pp. 187-211, 2002.

[7] G. Amitai, A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger, and S. Pietrokovski, "Network analysis of protein structures identifies functional residues," *J. Mol. Biol.*, vol. 344, pp. 1135-46, 2004.

[8] K. V. Brinda, and S. Vishveshwara, "Oligomeric protein structure networks: insights into protein-protein interactions," *BMC Bioinformatics*, vol. 6, p. 296, 2005.

[9] R. P. Chowdhury, M. S. Vijayabaskar, S. Vishveshwara, and D. Chatterji, "Molecular mechanism of in vitro oligomerization of Dps from Mycobacterium smegmatis: mutations of the residues identified by "interface cluster" analysis," *Biochemistry*, vol. 47, pp. 11110-7, 2008.

[10] A. Ghosh, and S. Vishveshwara, "A study of communication pathways in methionyl- tRNA synthetase by molecular dynamics simulations and structure network analysis," *Proc. Natl. Acad. Sci., U S A*, vol. 104, pp. 15711-6, 2007.

[11] N. Kannan, and S. Vishveshwara, "Aromatic clusters: a determinant of thermal stability of thermophilic proteins," *Protein Eng.*, vol. 13, pp. 753-61, 2000.

[12] A. Ghosh, and S. Vishveshwara, "Variations in clique and community patterns in protein structures during allosteric communication: investigation of dynamically equilibrated structures of methionyl tRNA synthetase complexes," *Biochemistry*, vol. 47, pp. 11398-407, 2008.

[13] B. Adamcsek, G. Palla, I. J. Farkas, I. Derenyi, and T. Vicsek, "CFinder: locating cliques and overlapping modules in biological networks," *Bioinformatics*, vol. 22, pp. 1021-3, 2006.

[14] K. V. Brinda, and S. Vishveshwara, "A network representation of protein structures: implications for protein stability," *Biophys. J.*, vol. 89, pp. 4159-70, 2005.

[15] K. V. Brinda, N. Mitra, A. Surolia, and S. Vishveshwara, "Determinants of quaternary association in legume lectins," *Protein Sci.*, vol. 13, pp. 1735-1749, 2004.

[16] A. Sethi, J. Eargle, A. A. Black, and Z. Luthey-Schulten, "Dynamical networks in tRNA:protein complexes," *Proc. Natl. Acad. Sci., U S A*, vol. 106, pp. 6620-5, 2009.

[17] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," *J. Mol. Graph.*, vol. 14, pp. 33-8, 27-8, 1996.

Received: November 24, 2010 Revised: December 10, 2010 Accepted: December 21, 2010