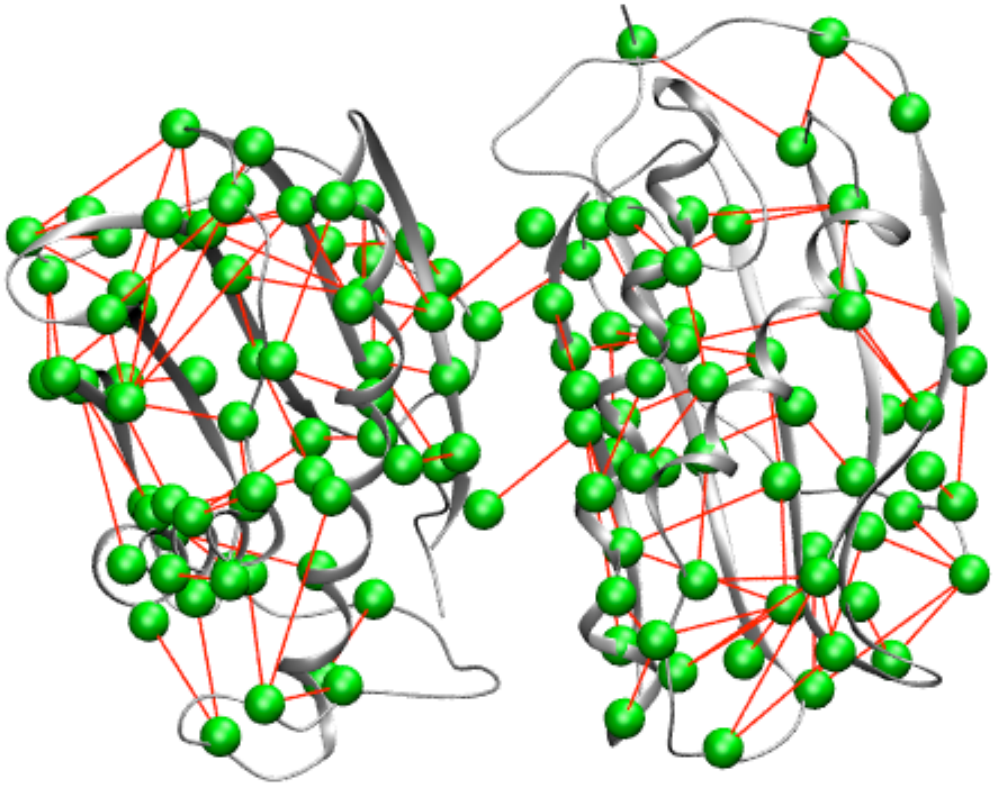
GraProStr – GRAPHS OF PROTEIN STRUCTURES

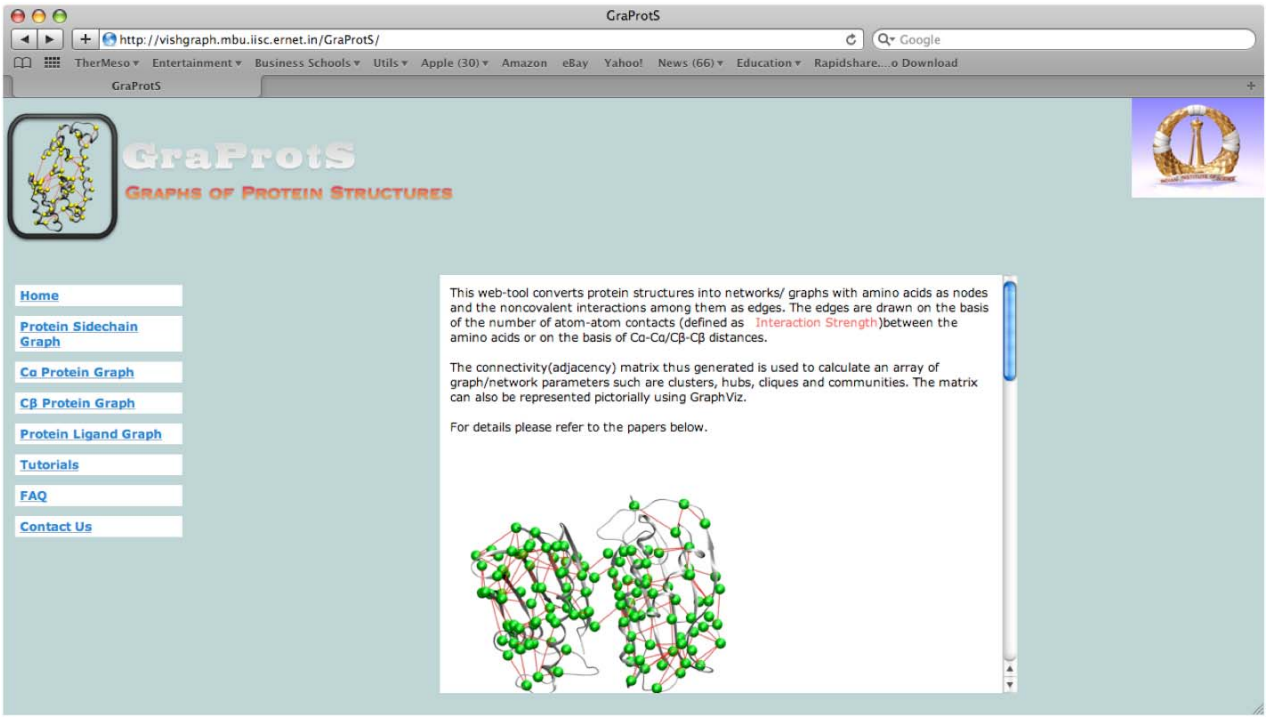


INTRODUCTORY TUTORIALS

Note: Please donot use Internet Explorer to view our web-tool.

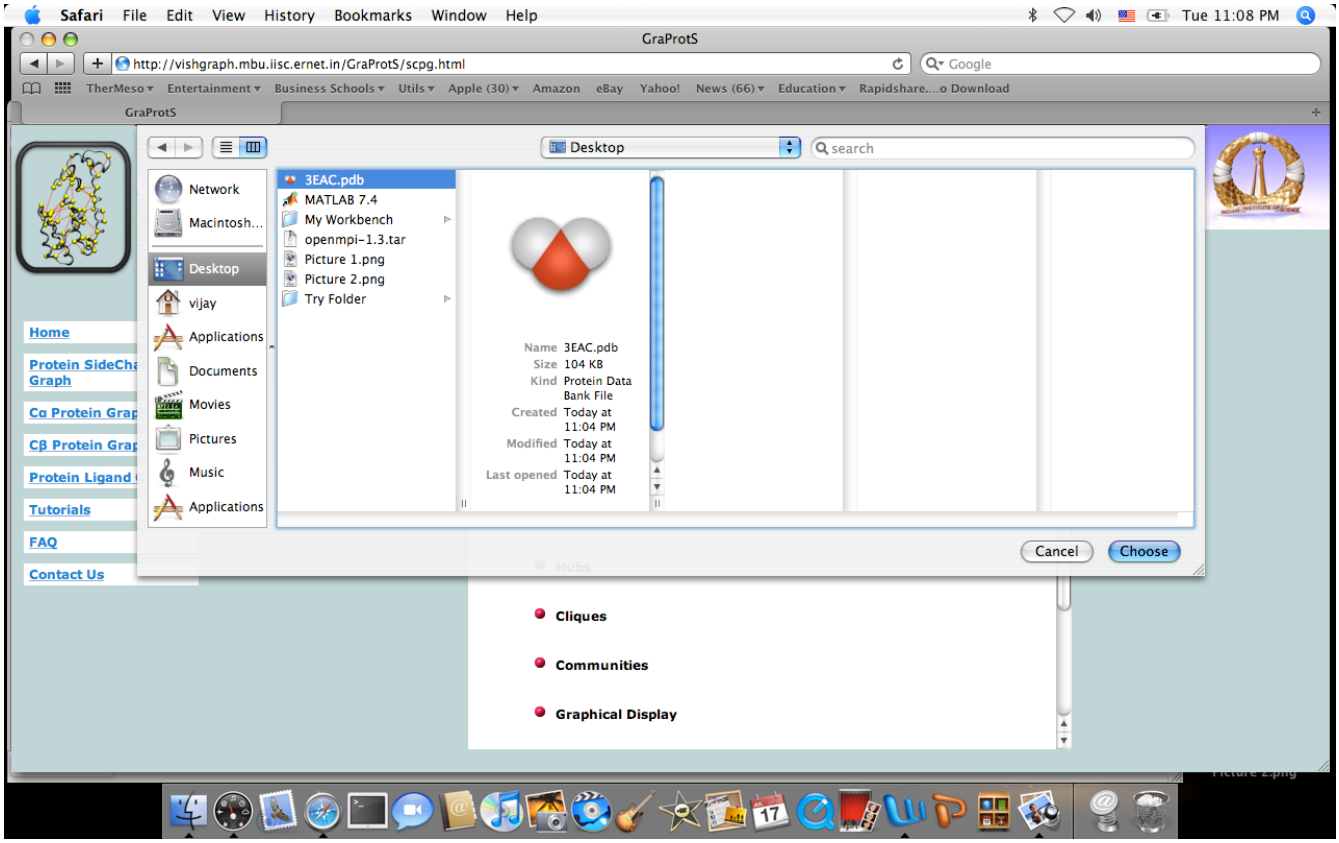
Open the Web-Tool

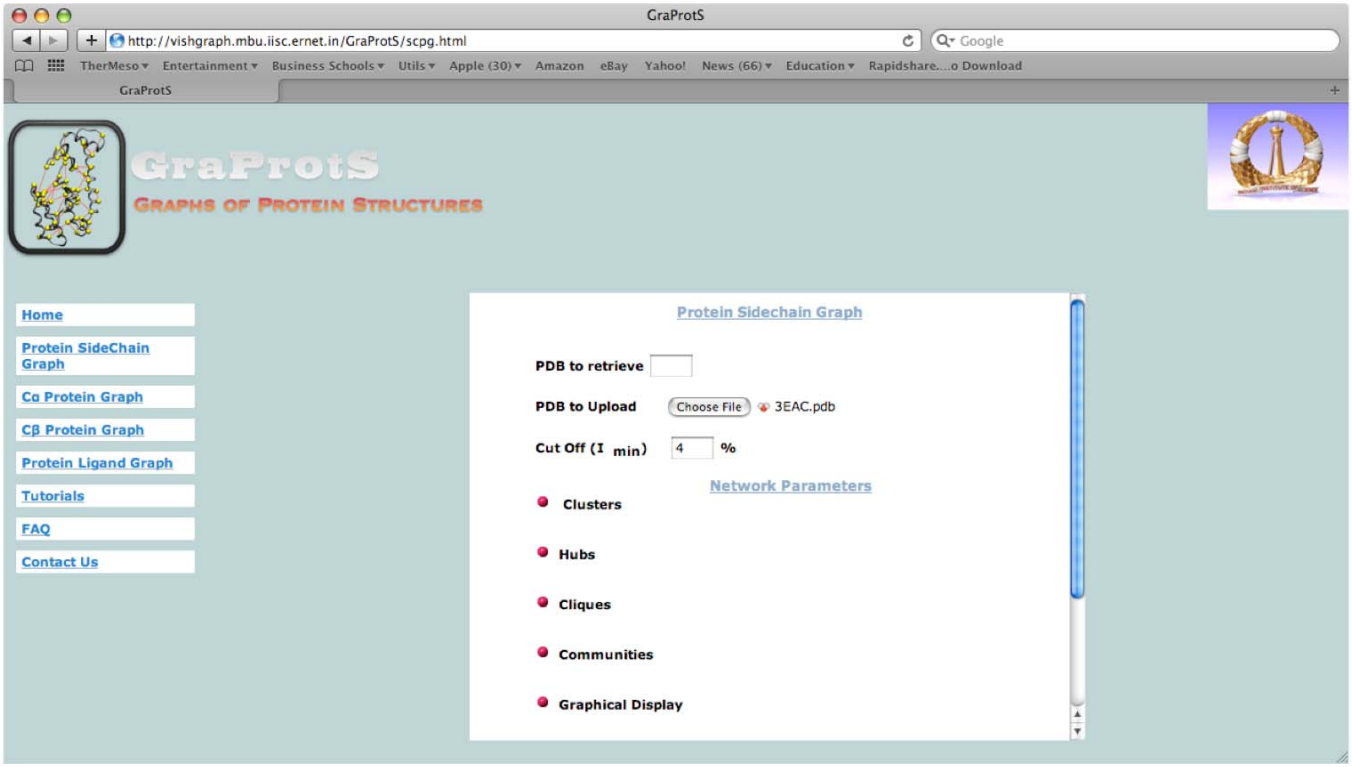
Type http://vishgraph.mbu.iisc.ernet.in/GraProtS/index.html in the address bar of the web browser. The left panel gives the different types of networks that can be generated and analyzed in this web-suite. Along with these modules are links to help the user to get started with or troubleshoot any errors if encountered.



Upload Protein Coordinates

The PDB is a Protein DataBank (http://www.rcsb.org) from which coordinates for the atoms in a macromolecule can be obtained. The coordinate files in PDB have a specific format and our web-tool is specifically follow this PDB format in reading the file that is uploaded.





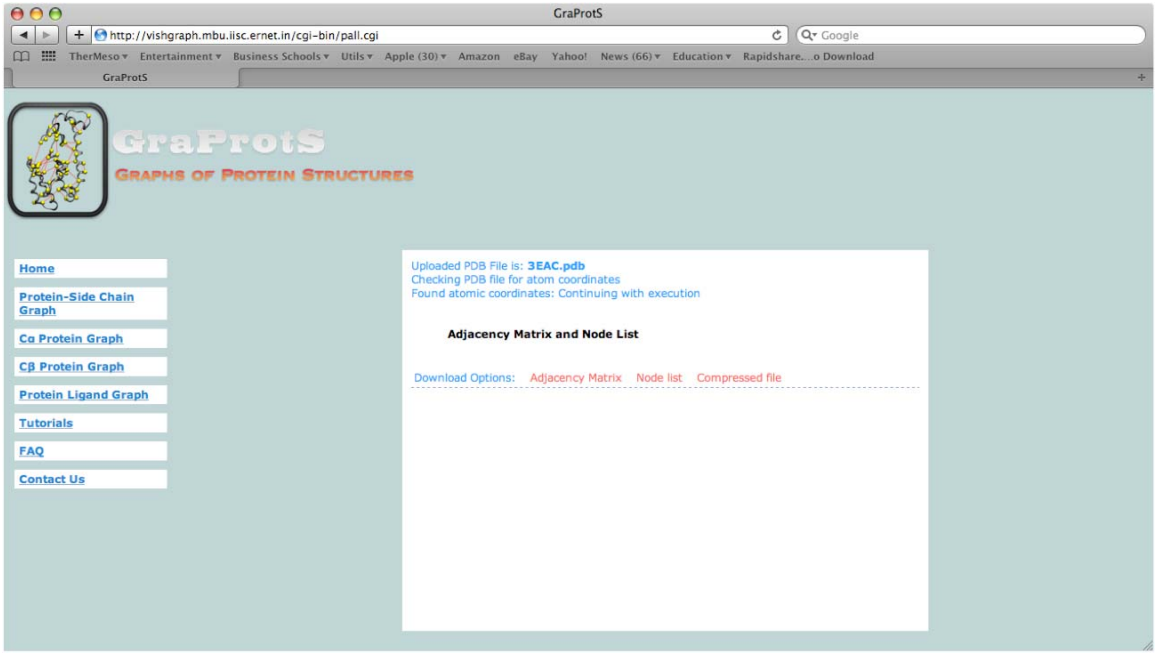
Instead you can also enter the 4 letter PDBID (3EAC in this example). This PDBID will be searched against the local PDB mirror site and then downloaded by the web-tool for further analysis.

Please note that the web-tool will give an error if the PDBID is does not match with the files in the mirror site. It also gives error if the mirror site is down.

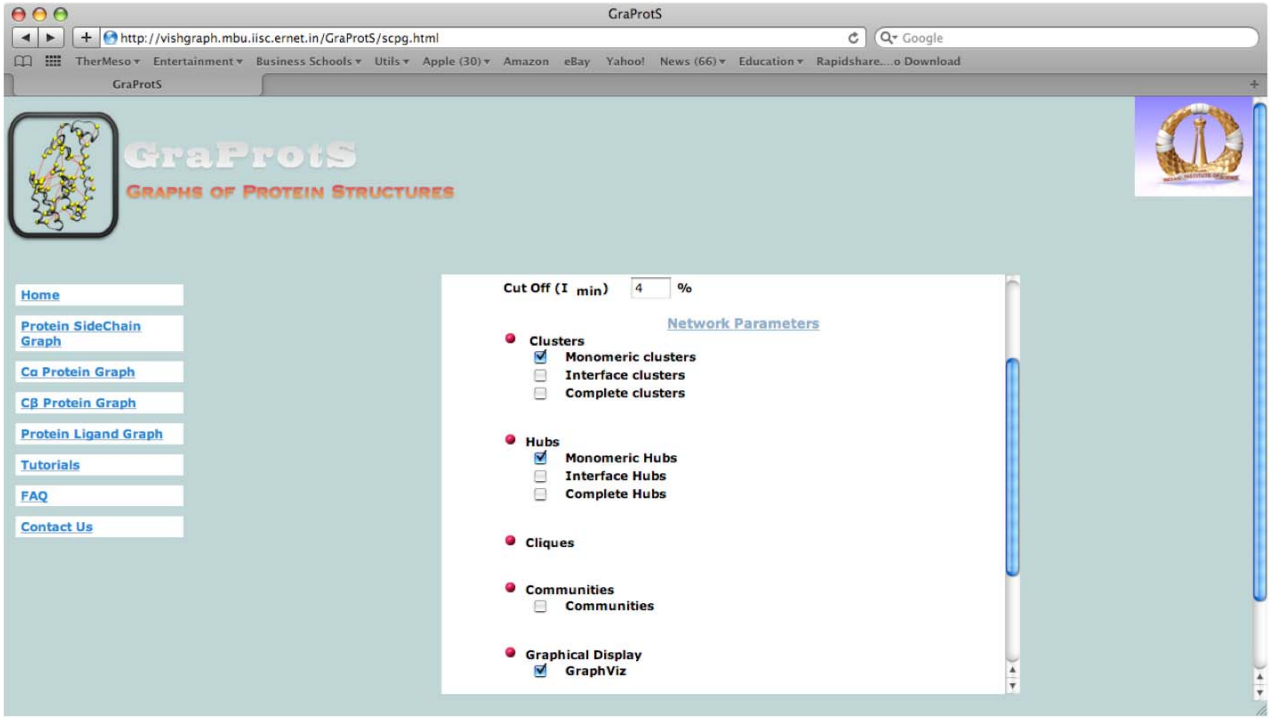
We have found a few anomalies in the mirror site with the RCSB site. So we strongly recommend that you download the site from RCSB and take a look at it before uploading.

Selecting Different Network Parameters

After uploading the pdb, you can select the parameters you want you analyze. If you do not select any parameters only the network will be generated. It will be given in the form of adjacency matrix and the corresponding nodes.



If you select a few parameters (Monomeric clusters, hubs and GraphViz), those parameters will be generated for the protein structure network.

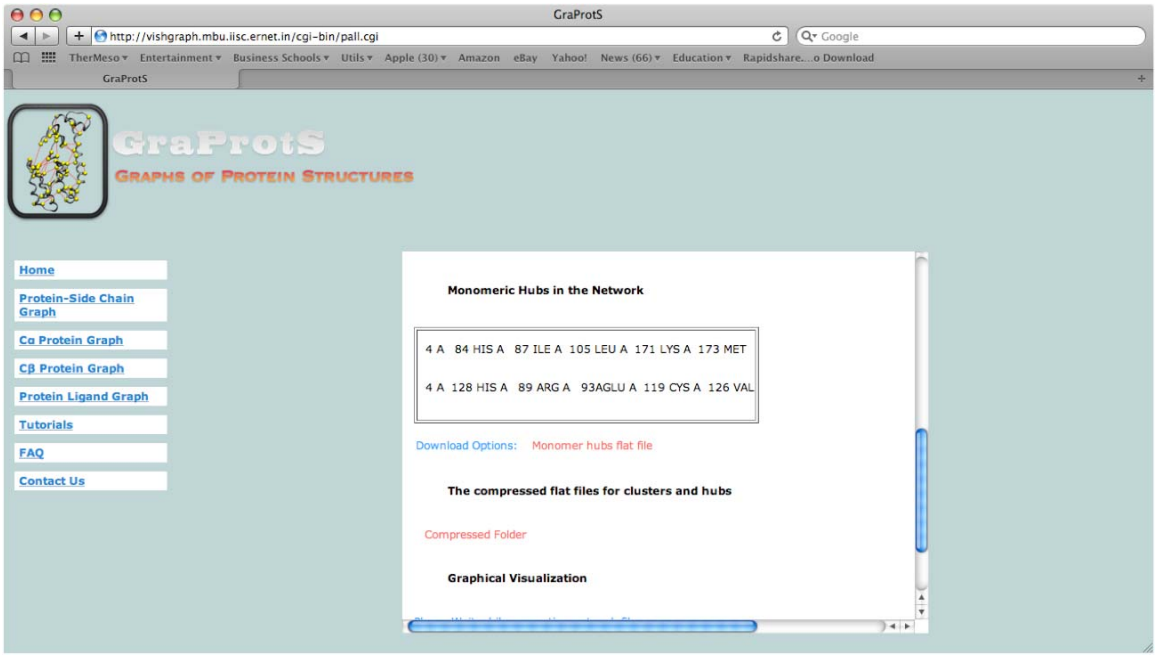


A default value is given as a cutoff for all the network modules (Here, cutoff = 4%). Feel free to play around with this value.

Running the Web-Tool

Press the SUBMIT button at the bottom of the page. If the program has read the pdb file correctly it will intimate you and then start generating the network parameters.

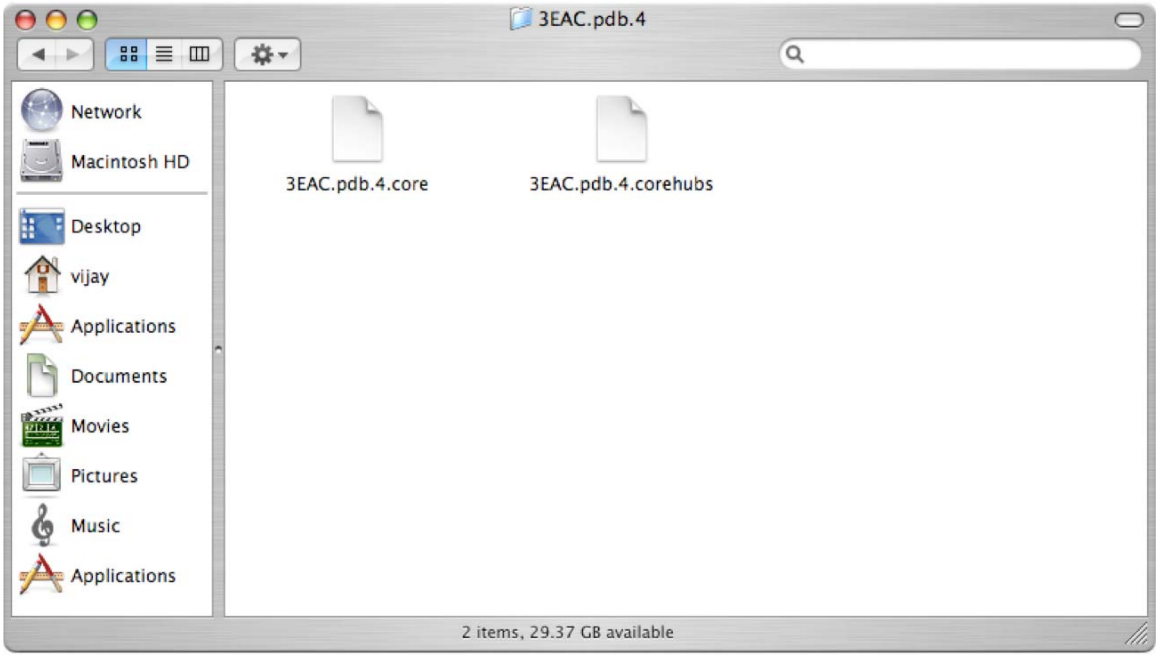
The clusters and hubs are displayed in the form of tables. Also, the flat file is given to the user for easy local storage and future use.



We have provided compressed folders for

1. Adjacency matrix and node list.
2. Clusters and Hubs
3. Cliques and Communities

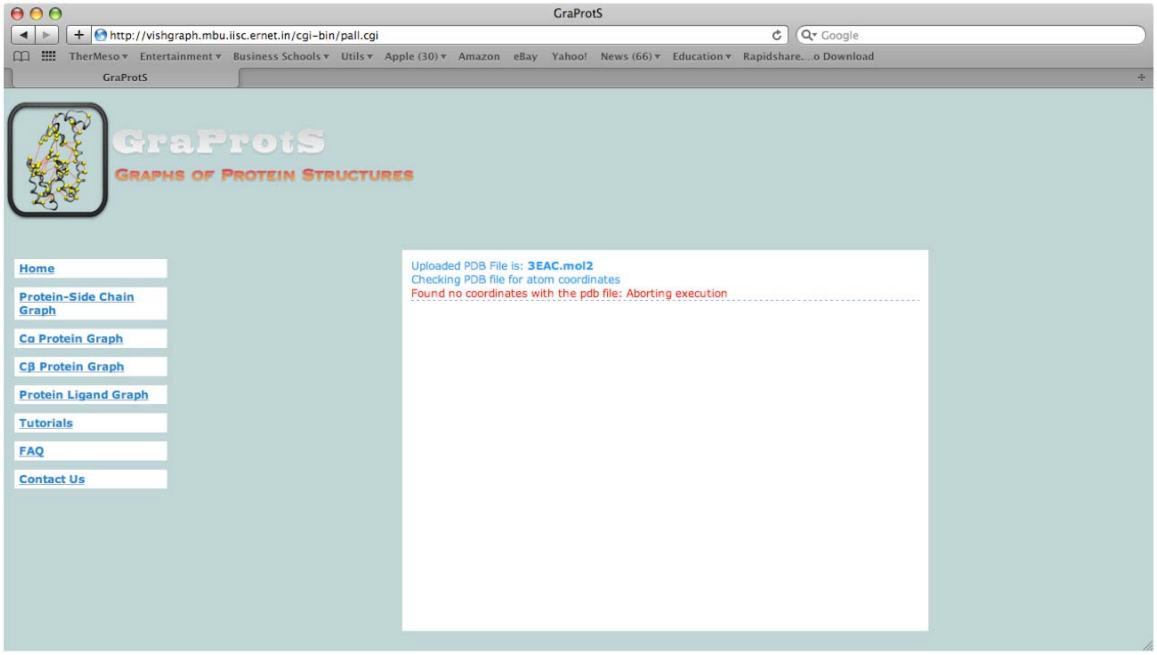
These folders enable convenient downloading and storage of network data.



The figure above shows the unzipped contents of the clusters and hubs file. The 3EAC.pdb.4.core gives the monomeric clusters of 3EAC at a cutoff of 4%.

In general “core” is used for denoting monomeric data, “inter” is used for denoting interface data. The word “core” does not mean the hydrophobic core of the protein molecule.

All the flat files from this web-tool can be viewed in any text processing softwares like (textpad, wordpad, textedit, vim etc).



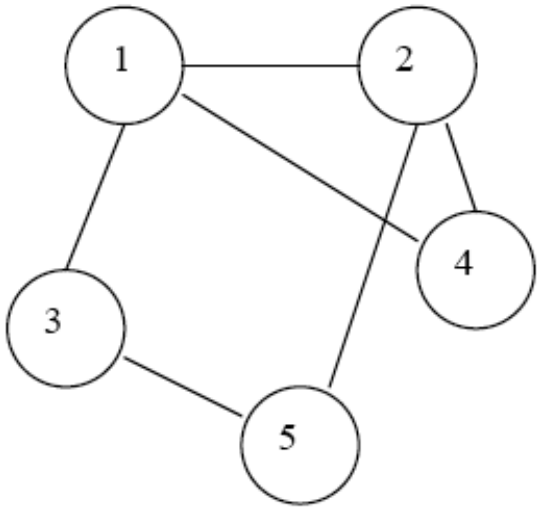
The above figure shows the error page if there are not coordinates in the file uploaded or the file is not of PDB format. For example we uploaded 3EAC.mol2 file, where the coordinates are given in mol2 format and not PDB format. The web-tool exits with an error in this case. Also, if you have entered the PDB ID (not uploaded the PDB file) and the tool returns an error, we recommend you to try to download the PDB and then upload it from your local computer. If the local PDB mirror site is down for maintenance the web-tool will return an error.

Analyzing the Results

1. Cut off, Adjacency Matrices and Node List

One of the two mandatory fields to be filled while using this web-tool is the cut off (the other field is the PDB file or PDB ID). The cut off is different for different Graphs/Networks. It is the Imin in Protein sidechain Graphs/Networks (PScN), Cα - Cα distance in Protein Cα Graphs/Networks and Cβ - Cβ distance in Protein Cβ Graphs/Networks.

Adjacency matrix is a mathematical representation of a graph. For eg.



The corresponding Adjacency matrix is a 5x5 matrix such as

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix}$$

where node 1 is connected to all the other nodes except 5 and itself. Therefore all the elements in row 1 except the 5^{th} element and 1^{st} element is 0.

Hence, Aij = 1 if there exists an edge between i and j. Aij=0 otherwise.

In a Protein Structure Graph/Network we construct an edge between any two amino acids i and j, if

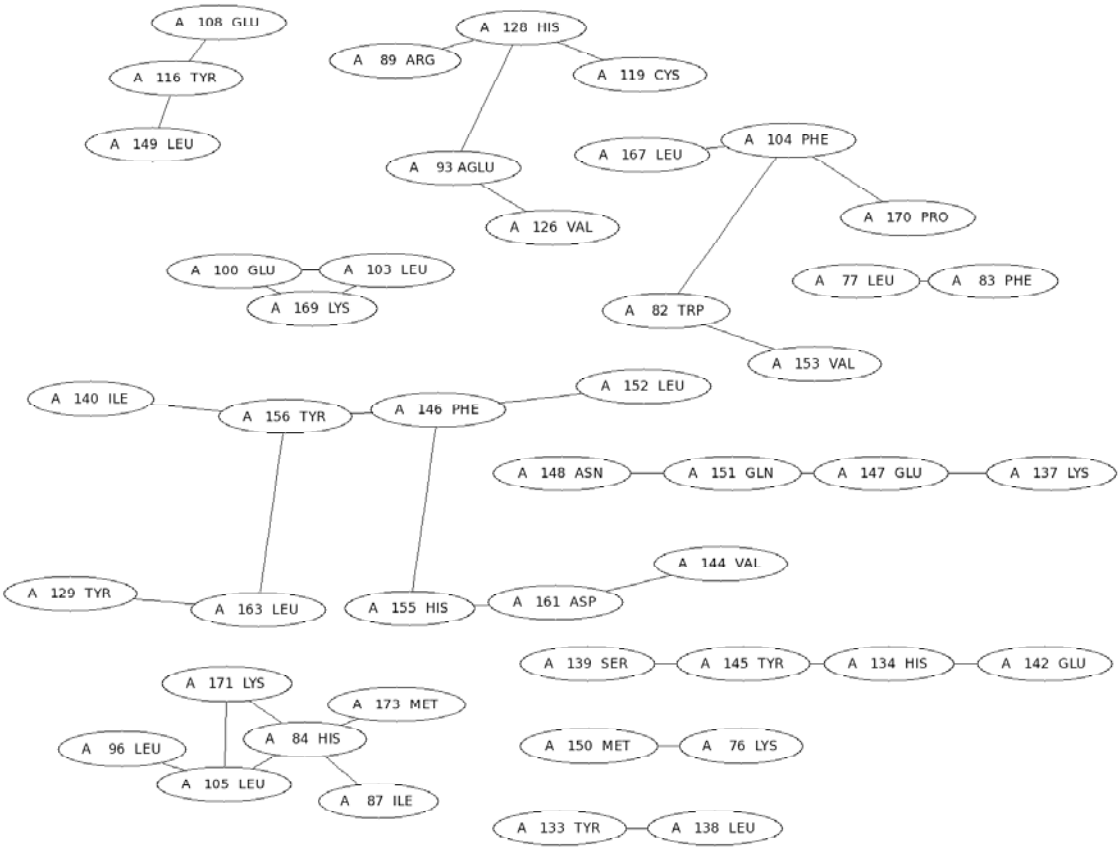
1. the interaction strength between them is greater than or equal to the cutoff value.
2. the Cα - Cα/ Cβ - Cβ distance between them is less than or equal to the cutoff value.

In general most of the interactions fall in the interaction strengths of 3% to 8%. The interactions range from 0% (any atom-atom contact) to as high as 20% (very high number of contacts only).

The Cα - Cα distance cutoff by default is given as 6.5 Å and Cβ - Cβ distance cutoff as 10 Å.

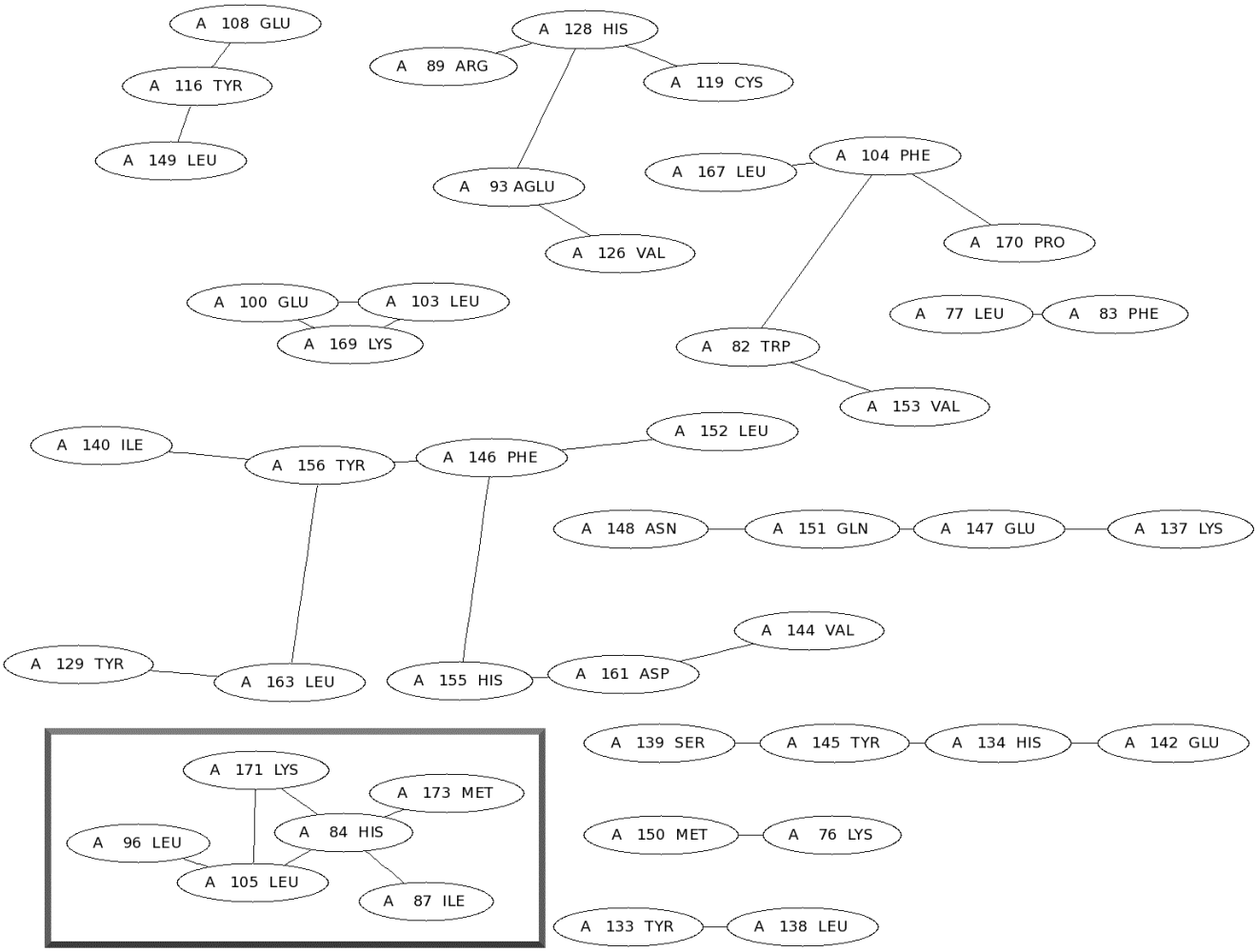
In the above-mentioned graph the nodes are annotated as 1, 2, 3, 4 and 5. In protein structure graphs/networks since each node is an amino acid, the nodes are identified as amino acids in the protein. Hence, for the protein structure graphs generated at a given cut off you will be provided with an adjacency matrix and a node list. Each element in the node list identifies to the corresponding row (node) in the adjacency matrix.

You can just use this adjacency matrix and node list to reconstruct the graph/network. We have used this information to generate a pictorial representation of the network (3EAC at 4% cut off in PScN) and it is given below.

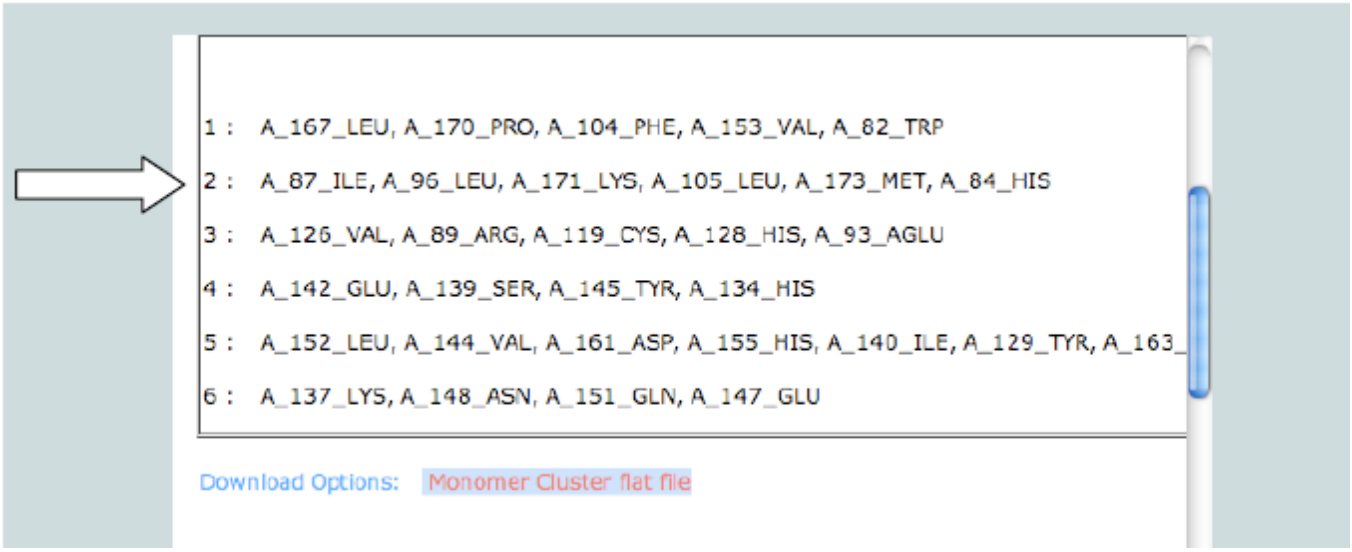


2. Clusters

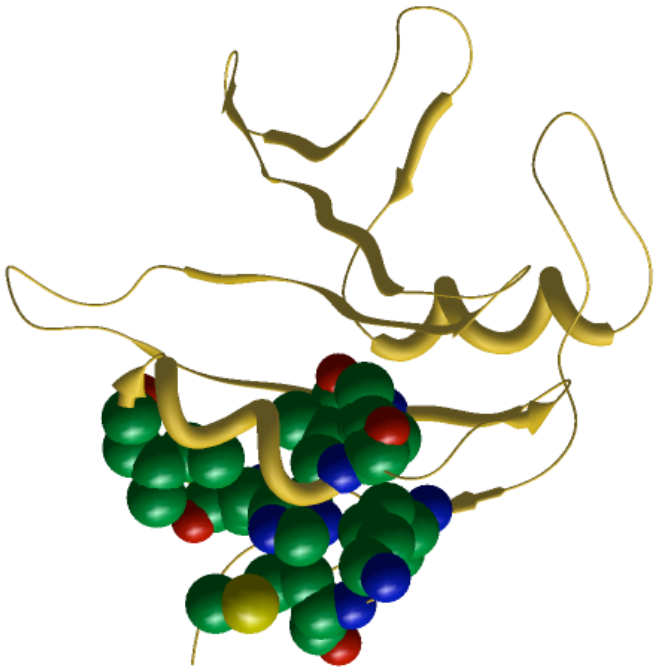
Clusters are groups of interacting residues. They are calculated using DFS algorithm. In the PScN network of 3EAC.pdb at cutoff=4% (figure below), we have highlighted a cluster of size 6 (total number of nodes = size of the cluster).



This highlighted cluster is being showed in the output (see cluster number 2) we obtained from the web-tool.



The cluster 2 is shown in the figure below.



3. Hubs

The hubs are highly connected nodes in the network. If the total number of edges incident on the node (called the degree of a node) is atleast 4 then we define that node as a hub.

In the network figure given above (and highlighted with a box) His84 is a hub since it has connections with Lys171, Leu105, Met173 and Ile87.

The format for analyzing the flat file is given below.

The first field of every line in the flatfile or the output is the degree of the hub node, followed by the hub with the <chain name> <resid> <res name>, followed by the connected nodes in the same format (<chain name> <resid> <resname>).

4. Cliques and Communities

A group of '*n*' nodes/residues are called as a k=*n* cliques, if each node is connected to every other node in the clique.

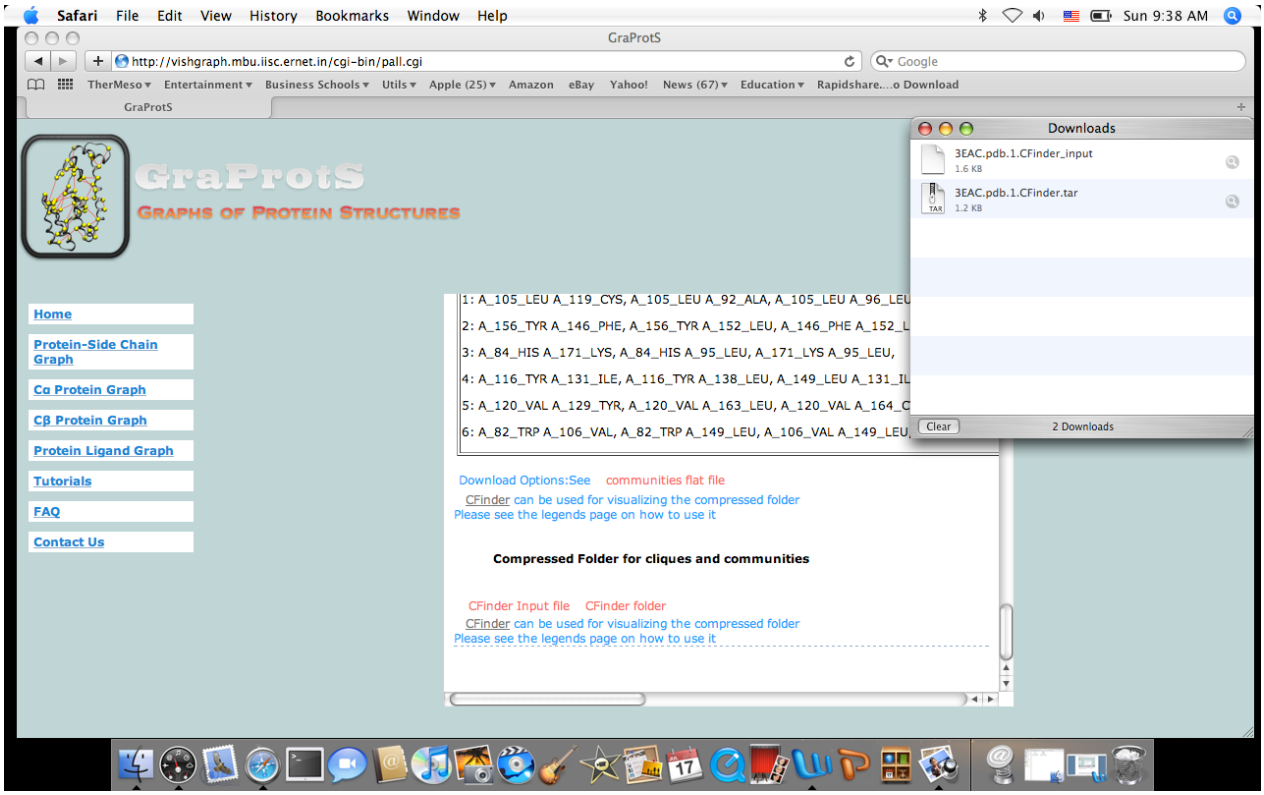
A community of k=*n* cliques is a collection of cliques that share n-1 nodes between them.

The cliques and communities are provided to you as a flat file. We have used CFinder to compute cliques and communities. The edges between nodes in a community are given separately. The information is provided in the format of – node1 node2, node2 node3, node1 node3, ….

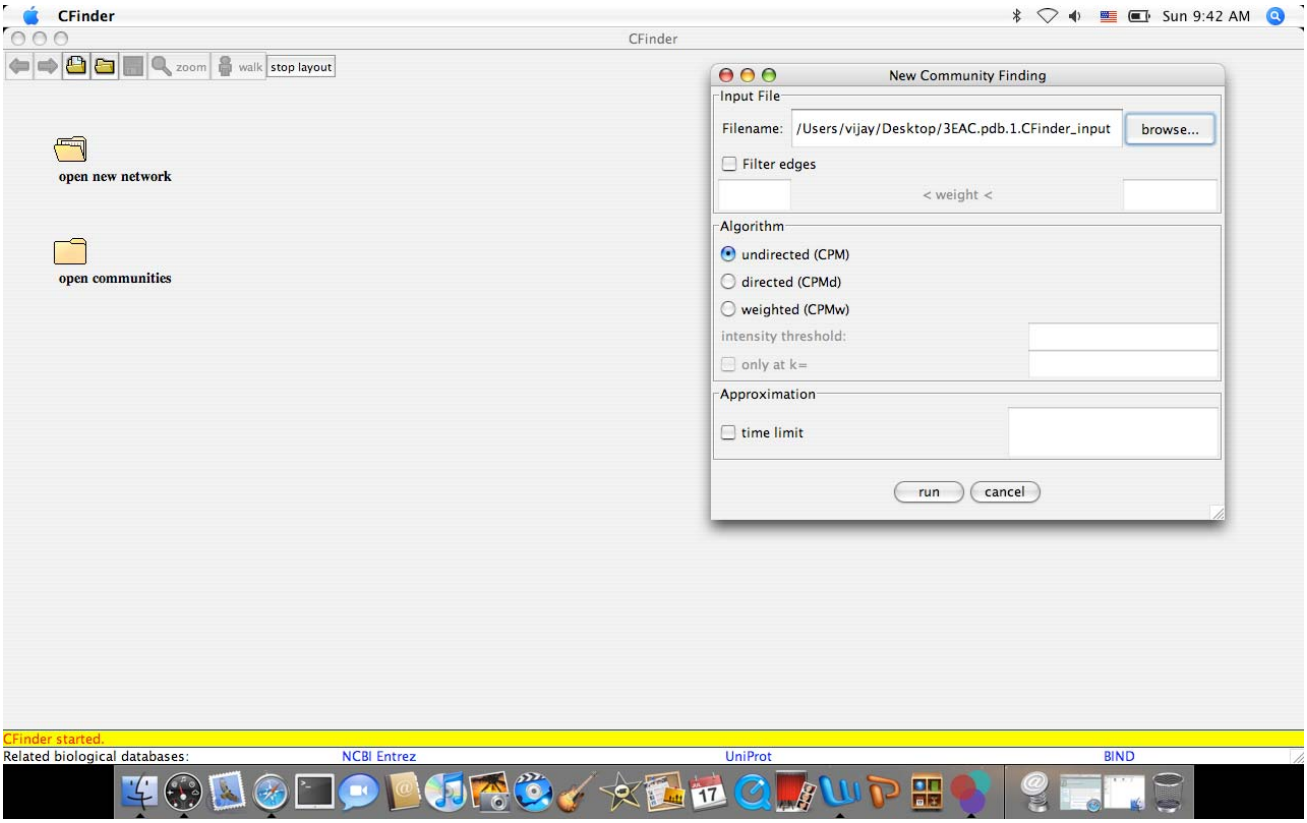
We have also provided the input folder for CFinder, so that you can use the pre-fabricated folder to re create the cliques and communities locally for better visualization and interpretation.

In the above example (3EAC.pdb, PScN with cutoff=1%) we have 15 cliques (k=3) and 6 communities of k=3 cliques. The edge information of the corresponding community is also given.

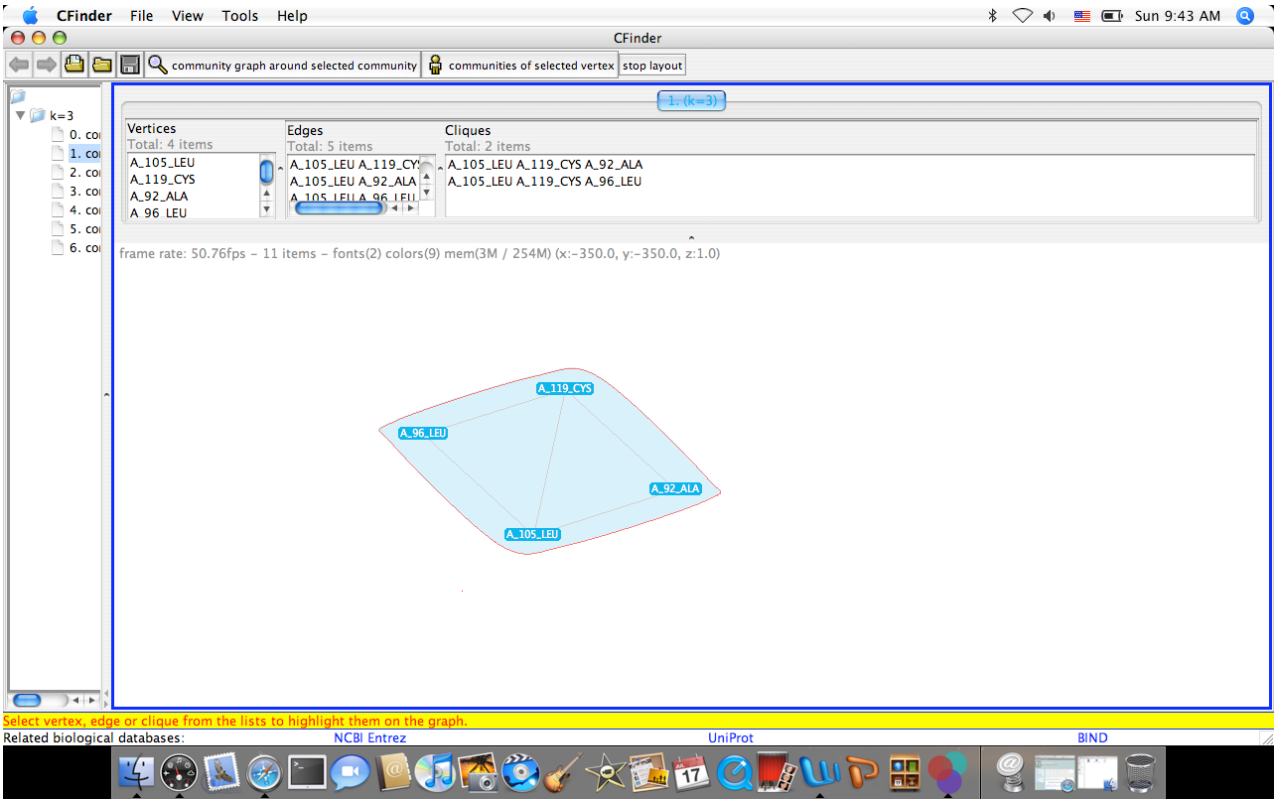
The compressed file can be downloaded, uncompressed and the files will give you all the information provided in the online output. To interpret the data in the files, please refer to the CFinder site.



You can download CFinder from this website (http://angel.elte.hu/cfinder/). After installing this software, you can use the CFinder input file (also given along with the compressed file), to load in CFinder.



Below, is given an example of a community of k=3 cliques (total cliques in this community is 2), (community number 2, when you run PScN for 3EAC.pdb, at cutoff = 1%), as seen from the output of CFinder.



5. Interface Properties

If you submit a multimer then the clusters, hubs, cliques can be divided as belonging to the individual monomers (within the same chain) and interface (between different chains). We have provided options for you to analyze either the monomeric properties or interface properties. In addition you can get the complete properties also which is the summation of both the monomeric and interface properties.

The references [3, 8, 10, 12-14] in the main text can help the readers understand the implications of the networks and their parameters in studying protein structures with greater clarity.