Systems Biology: A Personal View
VIII. Proteins as Networks: Centrality & Core-periphery

Sitabhra Sinha

IMSc Chennai

Can we say something about the important components of the protein using their contact networks ?

For this we can start by

Identifying the “central” nodes of the network

Centrality

Measures the importance of a node (or link) to the entire network

Wide variety of measures for vertex centrality:

1. Degree centrality or degree: number of links a node possesses

In many cases, nodes with the largest connections can be functionally critical – e.g., in spreading contagion

2. Eigenvector centrality: a node's importance is based on how many other important nodes it is connected to

Related measures are Katz centrality and PageRank (used by Google for its web-search algorithm)

3. Closeness centrality: measured in terms of mean geodesic distance of a node to other nodes

4. Betweenness centrality: how many times does a particular node occur along the shortest path between any pair of nodes

Eigenvector Centrality

Bonacich, 1987

In degree centrality, a node is scored in terms of the number of its neighbors
But all neighbors may not be equally important – e.g., a node connected to two hubs is more "important" than a node connected to two leaf nodes!

In eigenvector centrality each node is given a score proportional to the sum of the scores of its neighboring nodes

Let each node i be given a initial score $x_i(0)$ [e.g., = 1 for all i]
Starting from this initial guess, a better value of the centrality is calculated
$x_i(1) = \Sigma_j A_{ij} x_j(0)$ [using the defn of centrality as sum of neighbors centralities]
In matrix notation: $\mathbf{x}(1) = \mathbf{A}\,\mathbf{x}(0)$
Repeating this process iteratively for t steps, one gets $\mathbf{x}(t) = \mathbf{A}^t\,\mathbf{x}(0)$
Expressing $\mathbf{x}(0) = \Sigma_i c_i \mathbf{v}_i$ (i.e., a linear combination of the eigenvectors $\mathbf{v}_i$ of $\mathbf{A}$)
$\mathbf{x}(t) = \mathbf{A}^t \Sigma_i c_i \mathbf{v}_i = \Sigma_i c_i \lambda_i^t \mathbf{v}_i = \lambda_1^t \Sigma_i c_i [\lambda_i/\lambda_1]^t \mathbf{v}_i$
(where $\lambda_1 > \ldots > \lambda_i > \ldots > \lambda_N$ are the eigenvalues of $\mathbf{A}$)
As $\lambda_i/\lambda_1 < 1$, all terms other than the first decay as $t \to \infty \Rightarrow \mathbf{x}(t) \to c_1 \lambda_1^t \mathbf{v}_1$
Thus, centrality x satisfies $\mathbf{A}\,\mathbf{x} = \lambda_1\,\mathbf{x}$, i.e., it is proportional to the leading eigenvector of the adjacency matrix $\mathbf{A}$

Closeness Centrality

Measures how close a node is to other nodes of the network in terms of shortest paths

If d_{ij} is the length of a geodesic path from node i to node j,
the mean shortest path (avgd over all N nodes) from i to all other nodes in the network is $L_i=(1/N)\ \Sigma_j\ d_{ij}$
It is low for nodes that are separated from many other nodes only by short paths – and thus, e.g., communicates with the rest of the network faster
[Alternatively $L_i=(1/(N-1))\ \Sigma_j\ d_{ij}$ as d_{ii} can be taken to be zero]

The closeness centrality of a node i is the reciprocal of its avg distance (i.e., $C_i = 1/L_i$) from all other nodes of the network

If the network has multiple disconnected components, and i and j belong to different components, then d_{ij} is infinite
To resolve this problem closeness centrality can be defined in terms of harmonic mean of the distances between nodes: $C_i=(1/(N-1))\ \Sigma_{j(\neq i)}\ (1/d_{ij})$

doi:10.1016/j.jmb.2004.10.055 J. Mol. Biol. (2004) 344, 1135–1146

JMB

Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Network Analysis of Protein Structures Identifies Functional Residues

Gil Amitai, Arye Shemesh, Einat Sitbon, Maxim Shklar, Dvir Netanely
Ilya Venger and Shmuel Pietrokovski*

Use of Closeness Centrality to identify functionally important residues in a protein

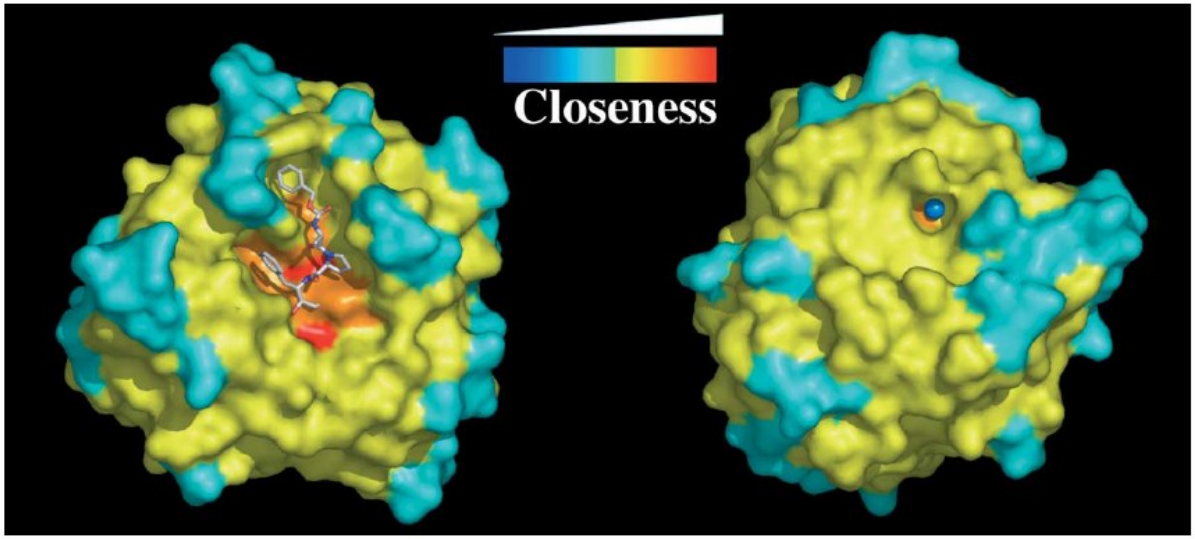
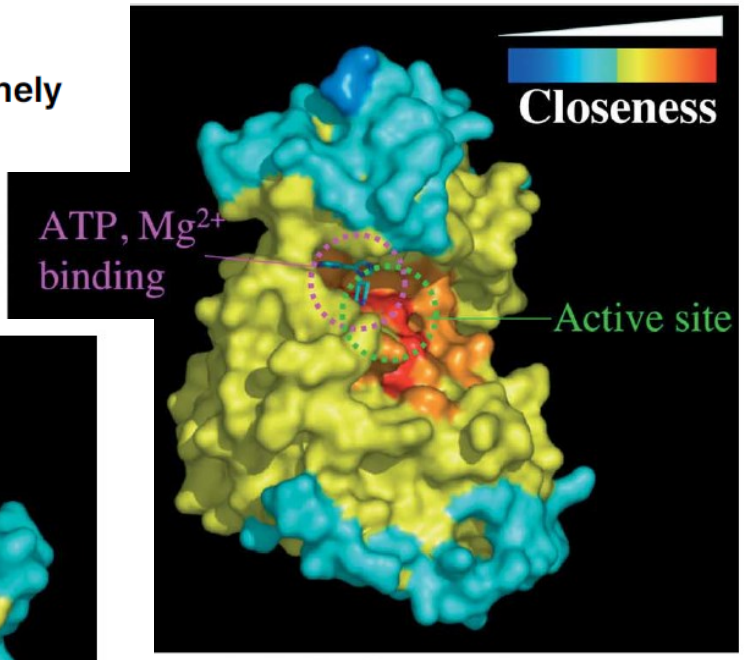


Figure 5. Closeness analysis of subtilisin DY protease. Closeness residue values are shown on the surface of the protein (PDB accession 1BH6). Closeness increases from blue to red. The left view shows the protease active site with a synthetic inhibitor, shown in sticks. The right view is related to the top by about 90° counterclockwise turn on the *Y* axis. It shows a Na atom in cation binding site B. Note the infrequency of residues with high closeness values and their exact overlap with the subtilisin active and cation binding sites.



Closeness centrality values of ERK2 MAP kinase. The active site and ATP-Mg2+ binding region have high closeness values.

Betweenness Centrality

Freeman, 1977

The importance of a node is measured in terms of how many geodesic paths in the network passes through it – nodes having high centrality of this kind will have large control over signals being sent by different nodes across the network

Consider the set of all geodesic paths in an undirected network in which there is at most one geodesic path between any pair of nodes
Betweenness centrality (BC_i) of a node i is the number of such paths that include i: $BC_i = \Sigma_{p,q} n^i_{pq}$,
where n^i_{pq} = 1 if node i is part of the geodesic path between p and q
n^i_{pq} = 0 otherwise

More generally, there can be more than one geodesic path between any pair of nodes – the standard extension is to give each such path between a pair of nodes i,j, a weight that is reciprocal of the total number of geodesic paths g_{ij} between the two nodes: $BC_i = \Sigma_{p,q} (n^i_{pq} / g_{pq})$

PHYSICAL REVIEW E, VOLUME 65, 061910

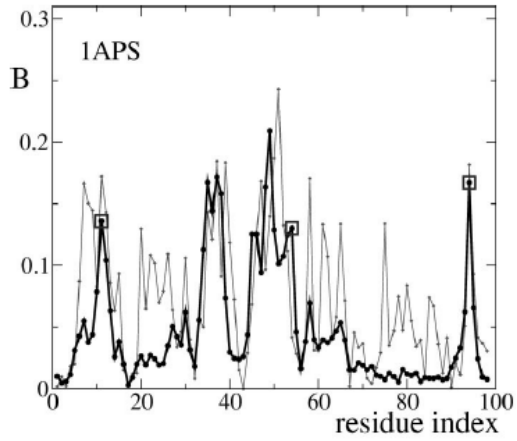
Small-world view of the amino acids that play a key role in protein folding

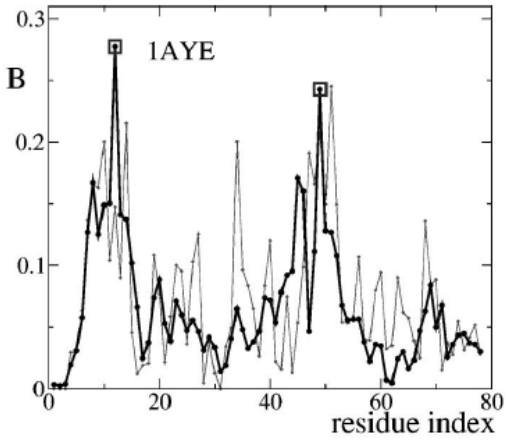
M. Vendruscolo,[1] N. V. Dokholyan,[2] E. Paci,[1,3] and M. Karplus[2,3]
[1]*Oxford Centre for Molecular Sciences, Central Chemistry Laboratory, South Parks Road, OX1 3QH Oxford, United Kingdom*
[2]*Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138*
[3]*Laboratoire de Chimie Biophysique, ISIS, Université Louis Pasteur, 4 rue Blaise Pascal, 67000 Strasbourg, France*
(Received 25 May 2001; revised manuscript received 21 March 2002; published 25 June 2002)

We use geometrical considerations to provide a different perspective on the fact that a few selected amino acids, the so-called "key residues," act as nucleation centers for protein folding. By constructing graphs corresponding to protein structures we show that they have the "small-world" feature of having a limited set of vertices with large connectivity. These vertices correspond to the key residues that play the role of "hubs" in the network of interactions that stabilize the structure of the transition state.

DOI: 10.1103/PhysRevE.65.061910 PACS number(s): 87.15.By, 64.60.Cn, 87.10.+e

Use of Betweenness Centrality to identify residues that contribute most to making the contact network "small-world"





Betweenness B in the transition state for proteins (thick lines). Nodes with large *B* are also usually key residues for forming the nucleus (squares). *B* values in native state (thin lines) shown for comparison.

"For the transition states of proteins … there is a small number (between 2 and 4) of residues (or regions) that have large *betweenness* values... Analysis of the transition states of these proteins have shown that there are certain residues, called *key residues*, which are critical for forming the nucleus that encodes the overall native structure… In all cases, they involve residues with large *betweenness*"

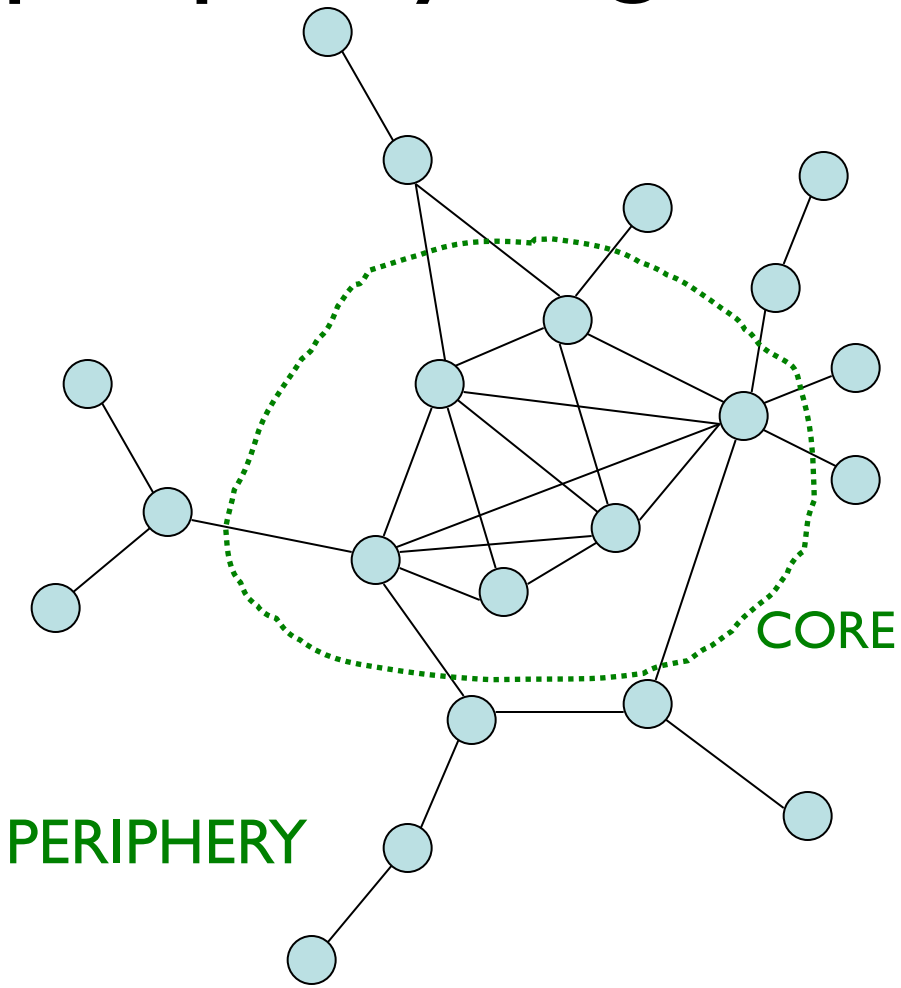
Can we say something about the important components of the protein using their contact networks ?

Identifying the network core of the proteins

Distinct from earlier notions of structural "core" as the set of residues which are completely inaccessible to solvent

The core may contain functionally critical residues !

Many networks possess a Core-periphery organization



K-Core Decomposition

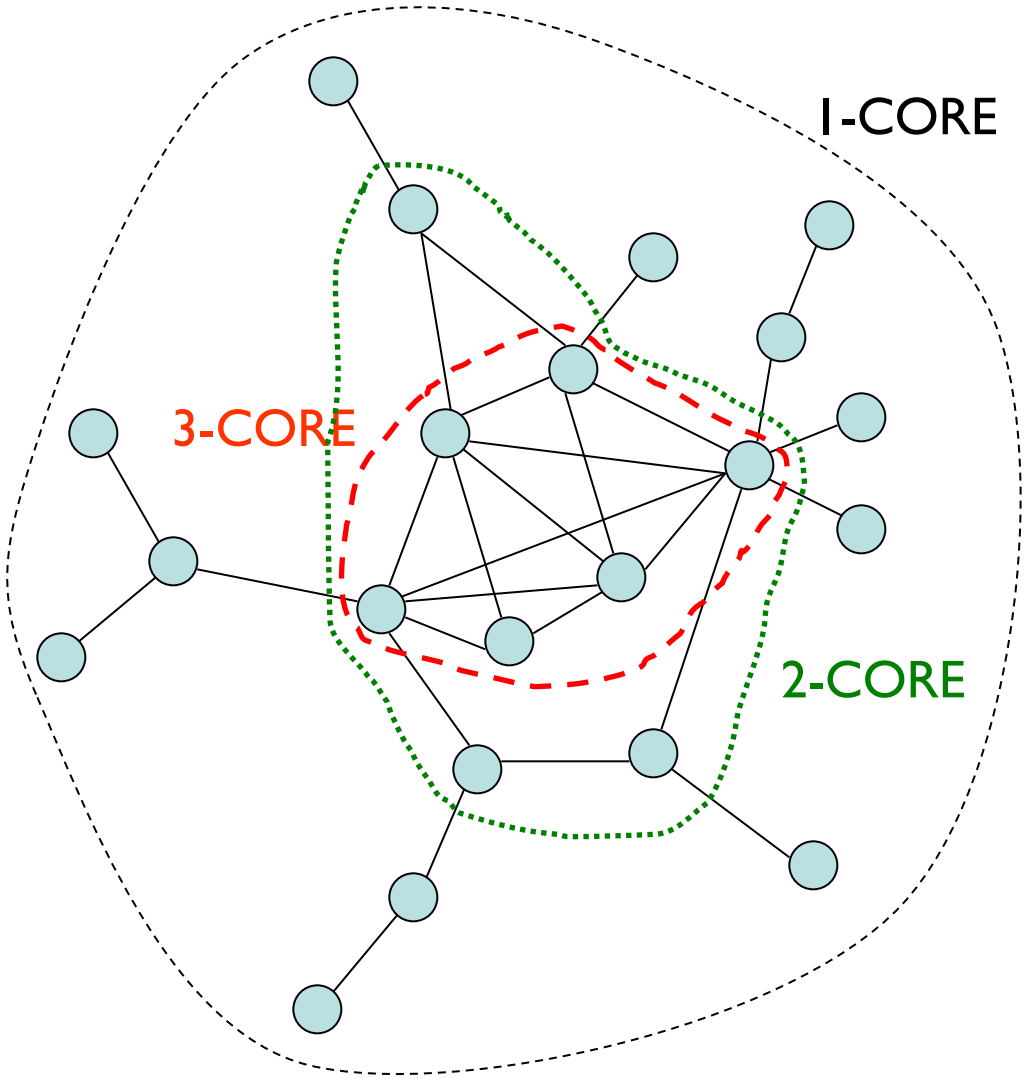
- Core decomposition, introduced by Seidman (1983), is a technique to obtain the fundamental structural organization of a complex network through a process of successive pruning
- Degree assortative networks show prominent core-periphery organization
- The k-core decomposition was recently applied to a number of real-world networks:
  - the Internet(Alvarez et al, 2005)
  - WWW (Kirkpatrick et al, 2005),
  - neuronal network of *C. elegans* (Chatterjee & Sinha, 2007) etc.
- The most efficient spreaders are those located within the core of the network (Kitsak et al, 2010)

K-Core Decomposition

- Defn: The *k-core* of a network is the subnetwork containing all nodes that have degree *at least* equal to *k*.
- An iterative procedure for determining the *k*-core is
  - (i) to remove all nodes having degree less than *k*,
  - (ii) check the resulting network to see if any of the remaining nodes now have degree less than *k* as a result of (i), and if so
  - (iii) repeat steps (i)-(ii) until all remaining nodes have degree at least equal to *k*.
- This resulting network is the *k*-core of the original network.
- In particular, the 2-core of a network is obtained by eliminating all nodes that do not form part of a loop (a closed path through a subset of the connected nodes).
- There exist at least *k* paths between any pair of nodes belonging to a *k*-core.

Seidman (1983)

k-Core Decomposition



Example: K-Core Decomposition of a Protein

PDB ID : 3JS3 A (3-dehydroquinate dehydratase)

1-Core = 253 Residues

Figure created by Arnold Emerson

1-Core = 253 Nodes

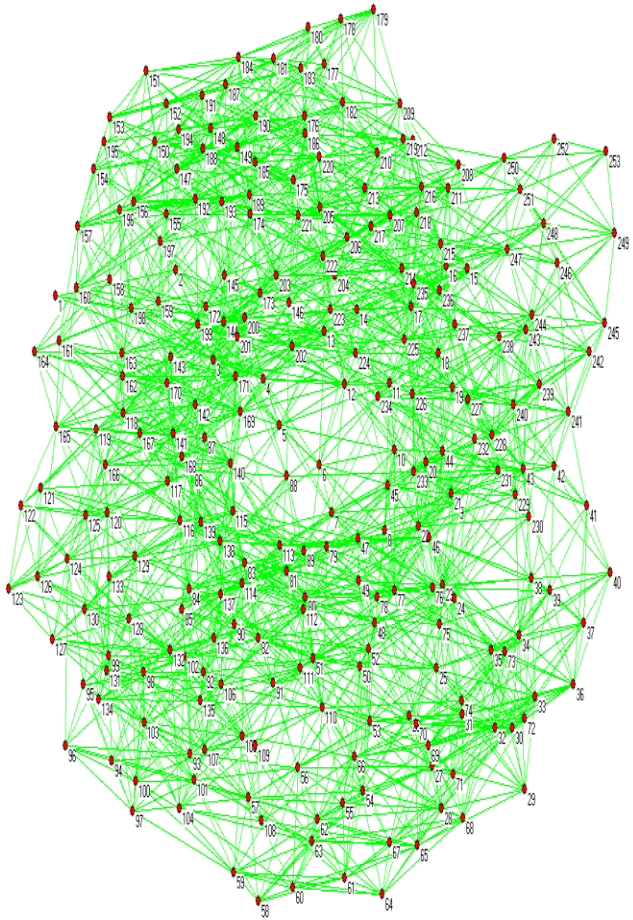


Figure created by Arnold Emerson

2-Core = 253 Residues

Figure created by Arnold Emerson

2-Core = 253 Nodes

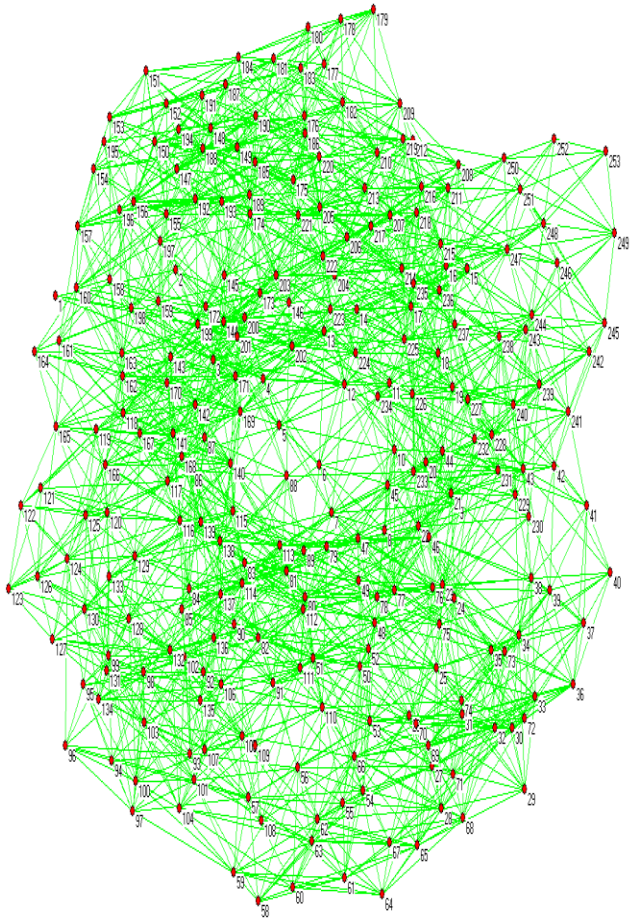


Figure created by Arnold Emerson

3-Core = 253 Residues

Figure created by Arnold Emerson

3-Core = 253 Nodes

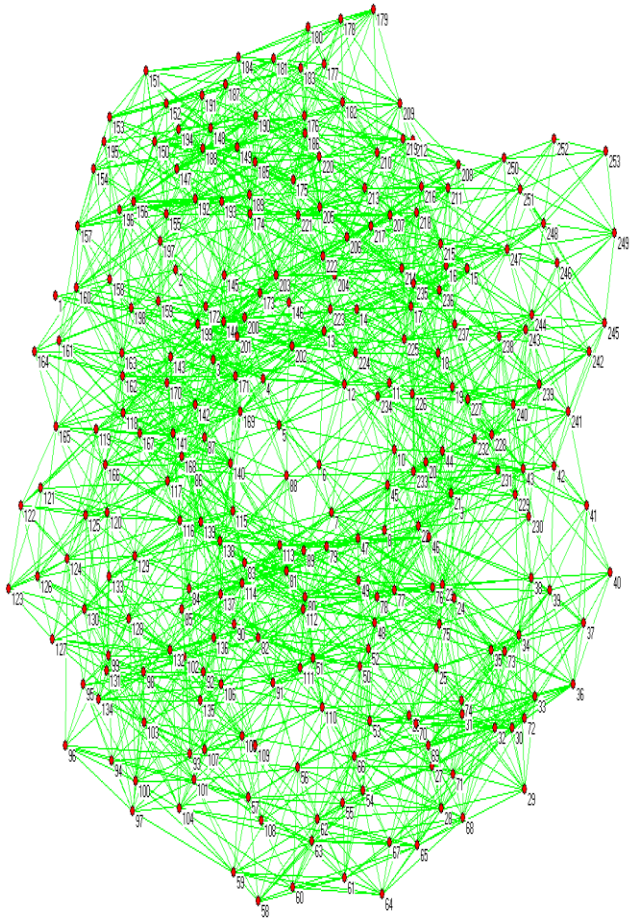


Figure created by Arnold Emerson

4-Core = 253 Residues

Figure created by Arnold Emerson

4-Core = 253 Nodes

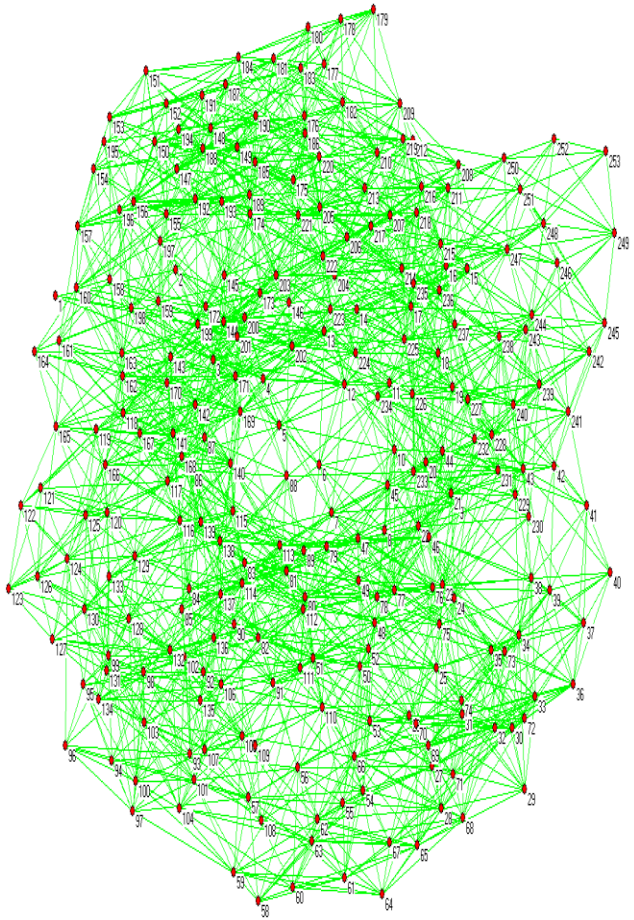


Figure created by Arnold Emerson

5-Core = 253 Residues

Figure created by Arnold Emerson

5-Core = 253 Nodes

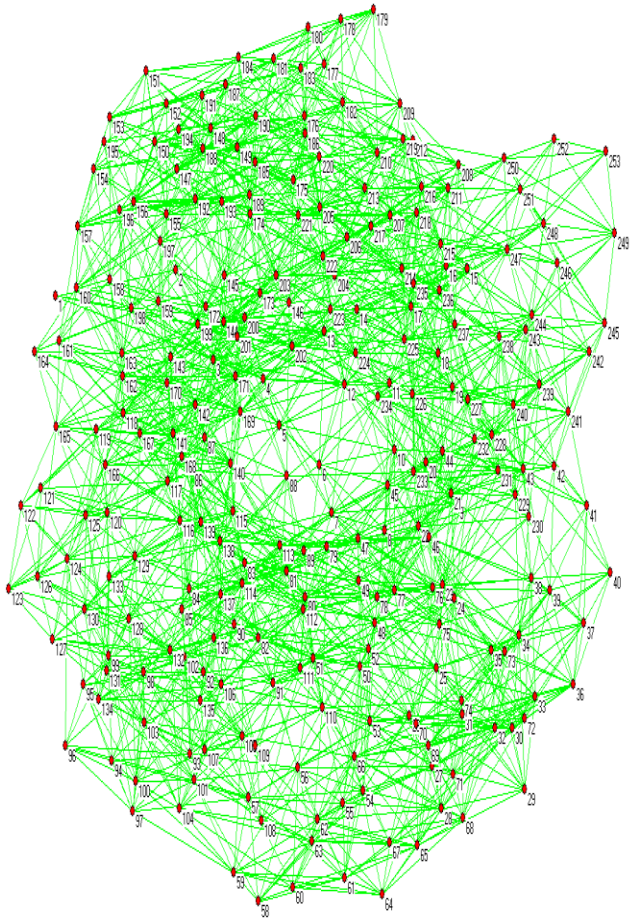


Figure created by Arnold Emerson

6-Core = 252 Residues

Figure created by Arnold Emerson

6-Core = 252 Nodes

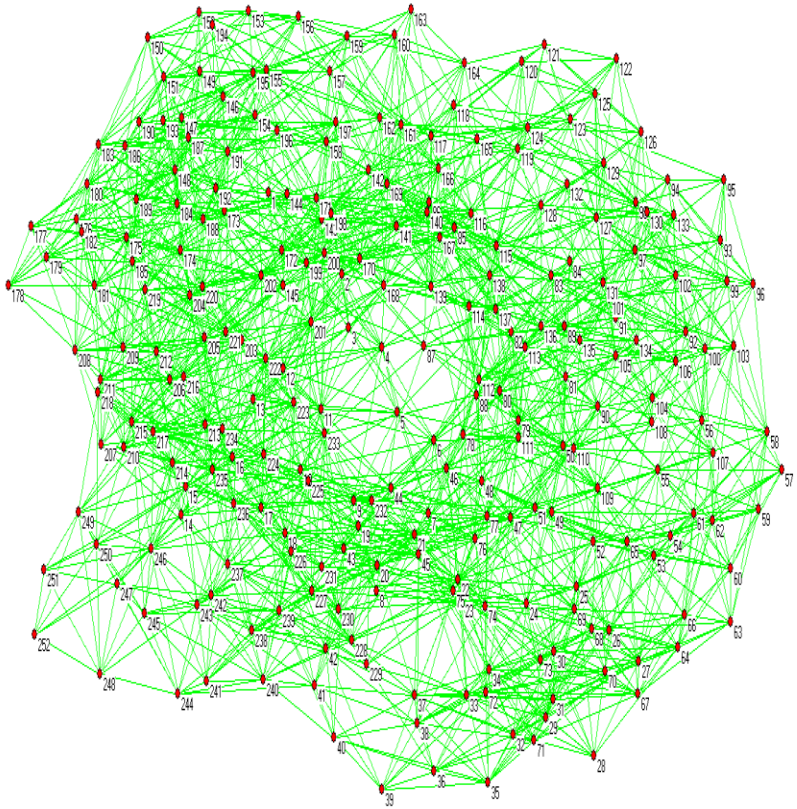


Figure created by Arnold Emerson

7-Core = 250 Residues

Figure created by Arnold Emerson

7-Core = 250 Nodes

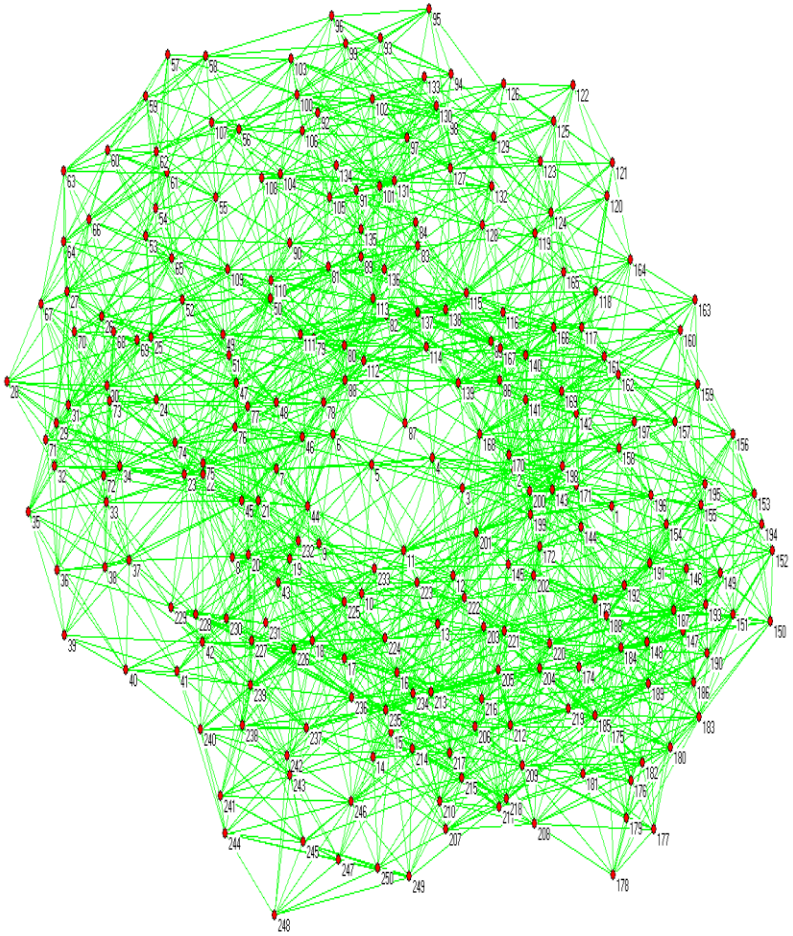


Figure created by Arnold Emerson

8-Core = 250 Residues

Figure created by Arnold Emerson

8-Core = 250 Nodes

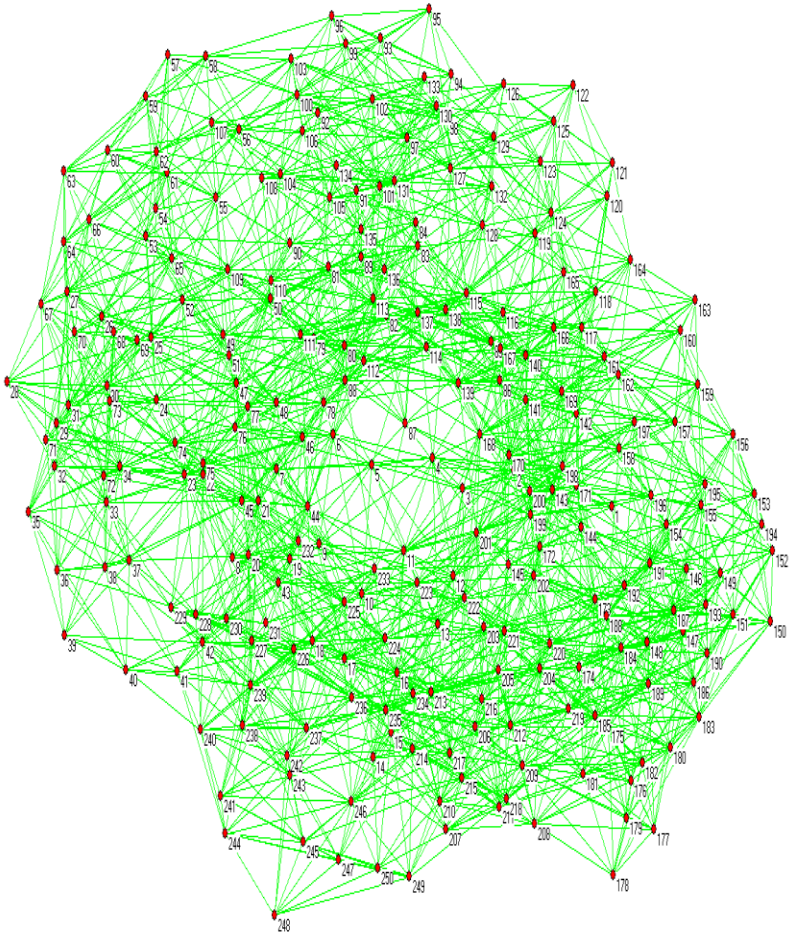


Figure created by Arnold Emerson

9-Core = 248 Residues

Figure created by Arnold Emerson

9-Core = 248 Nodes

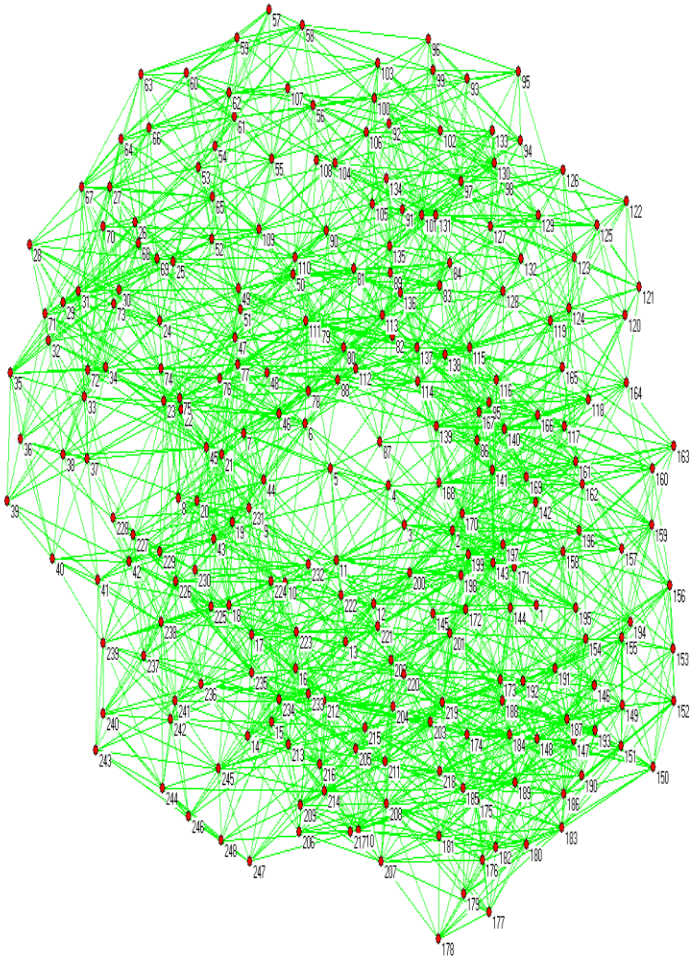


Figure created by Arnold Emerson

10-Core = 240 Residues

Figure created by Arnold Emerson

10-Core = 240 Nodes

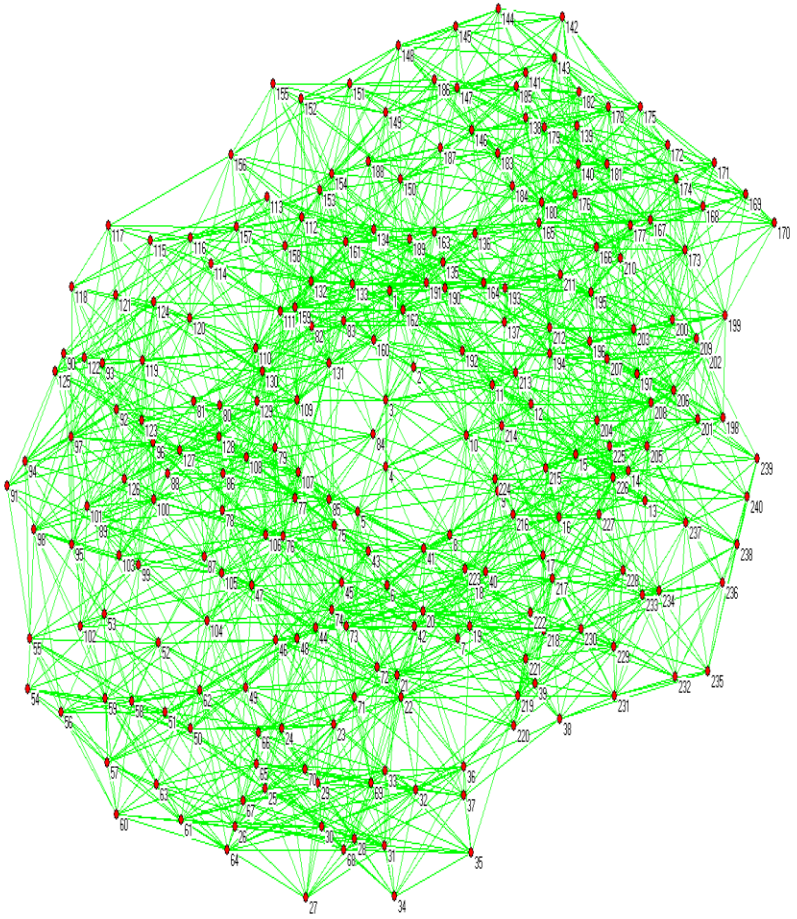


Figure created by Arnold Emerson

11-Core = 177 Residues

Figure created by Arnold Emerson

11-Core = 177 Nodes

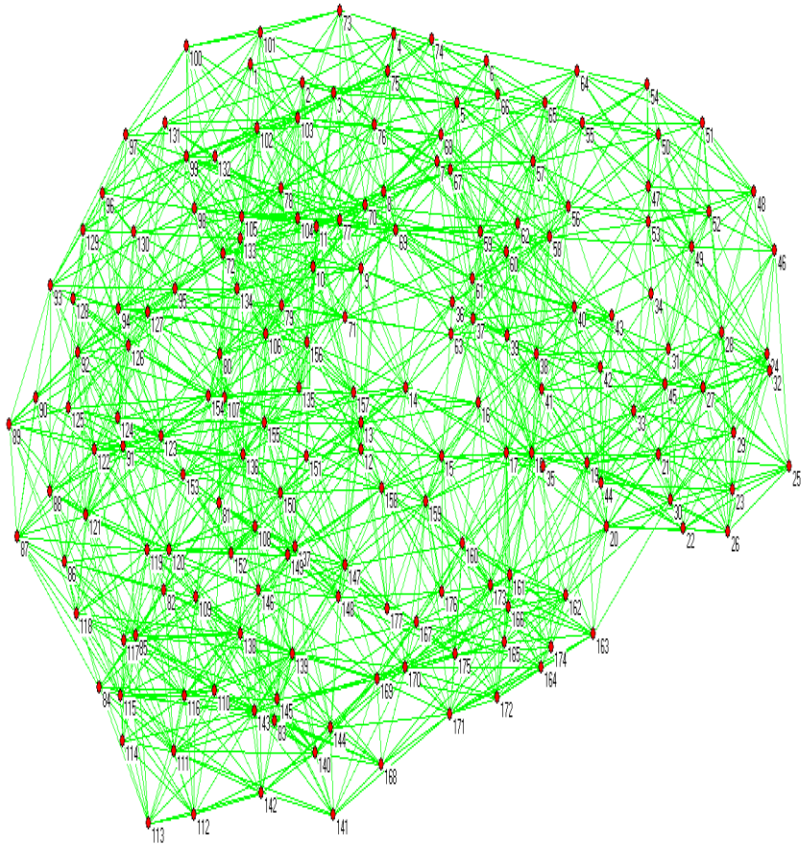


Figure created by Arnold Emerson

12-Core = 112 Residues

Figure created by Arnold Emerson

12-Core = 112 Nodes

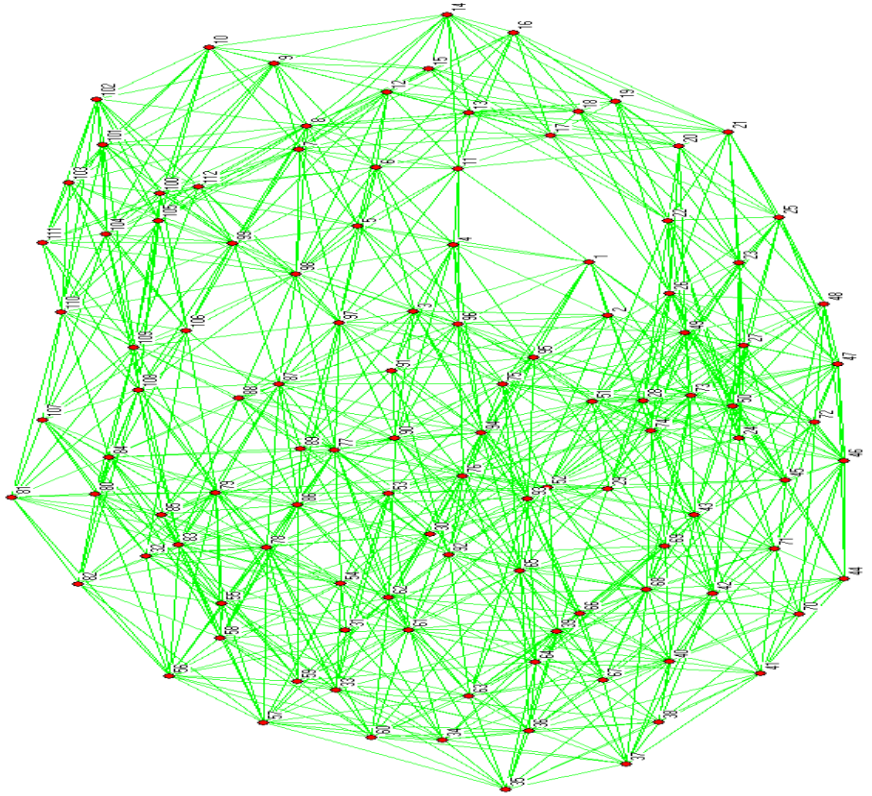


Figure created by Arnold Emerson

13-Core = 0 Residues

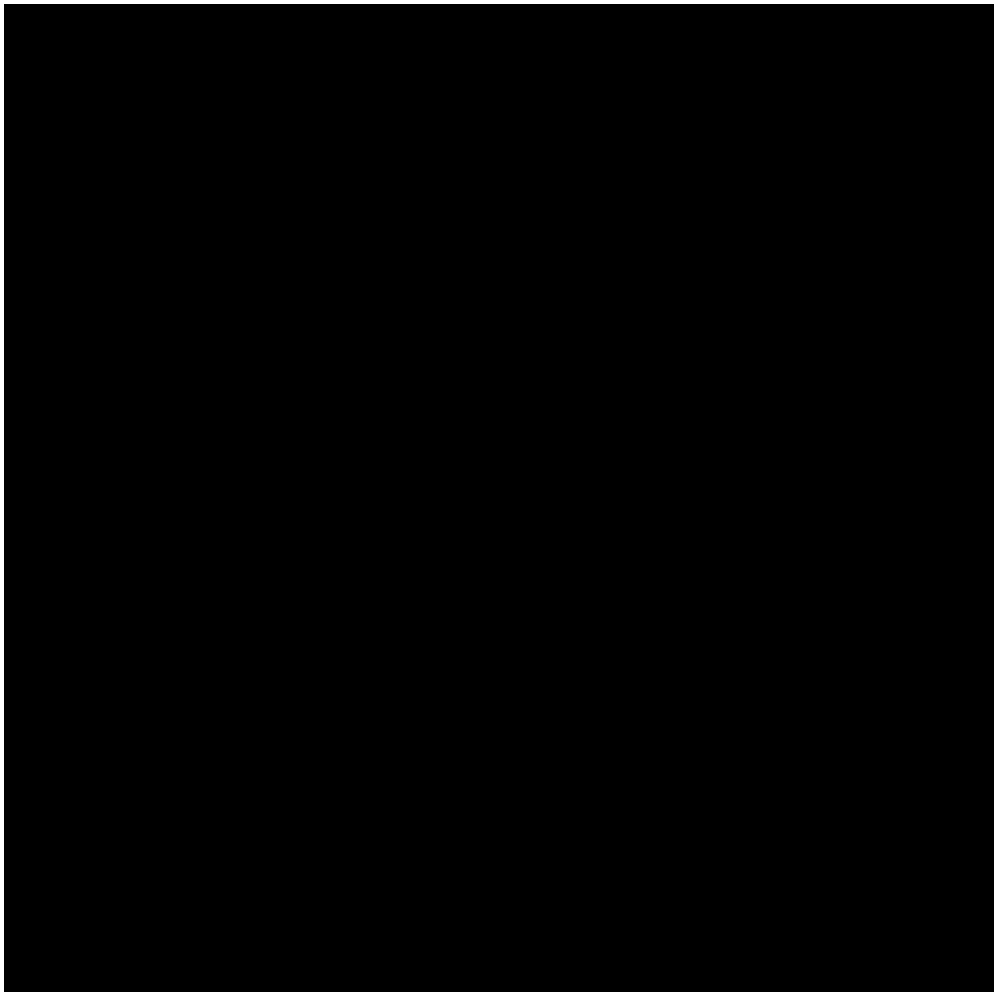
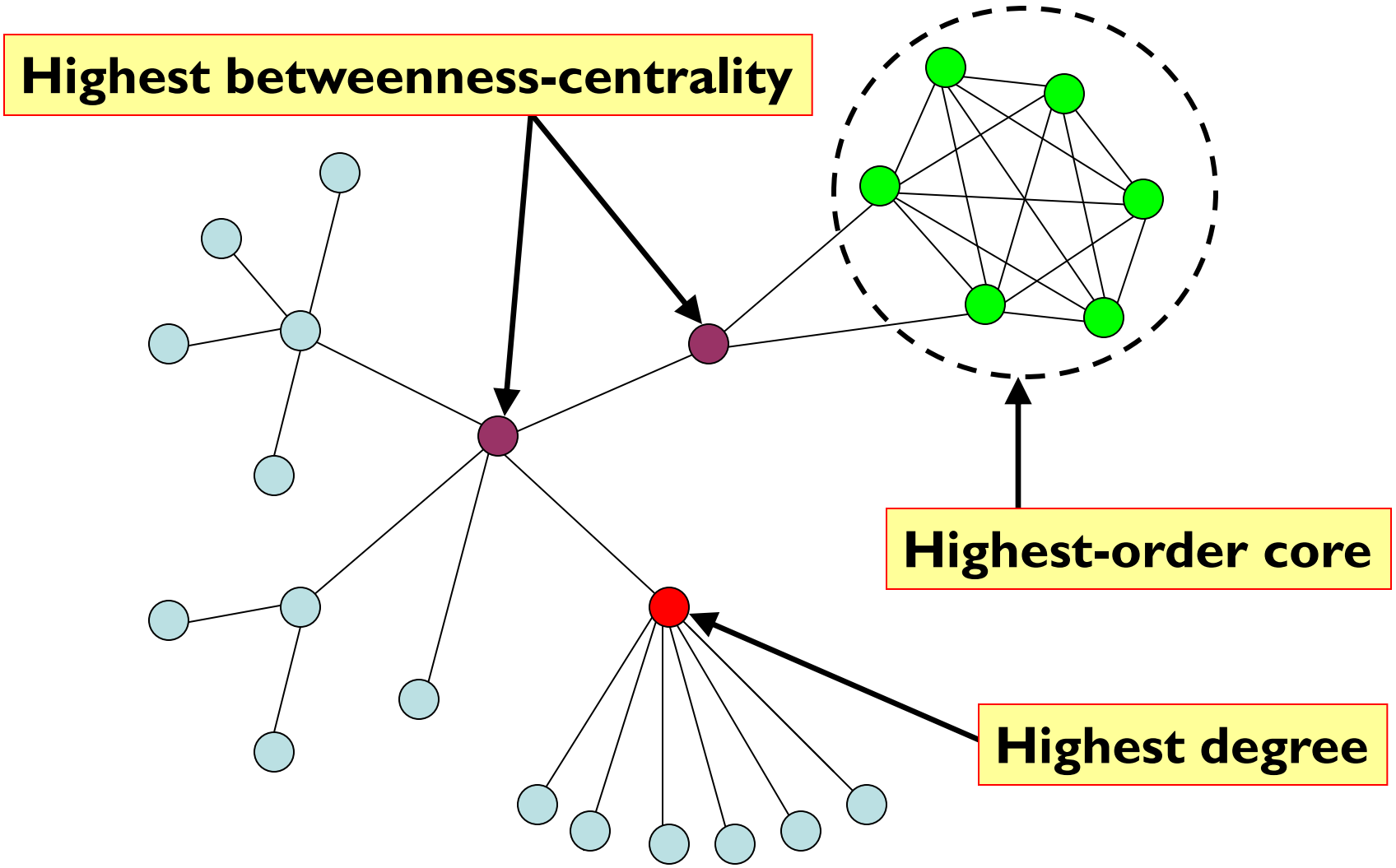


Figure created by Arnold Emerson

13-Core = 0 Nodes

Figure created by Arnold Emerson

How is core order membership distinct from other node-specific measures ?



Questions

- What is special about residues belonging to the inner core of a protein ?
- Could they be functionally important ?
- How to check this hypothesis ?

Functional Importance of inner core residues

- **Solvent Accessibility**
  - Provides information about whether amino acid residues in proteins with known structures are accessible to solvent

  Inner core residues have lower accessibility than those at the periphery

- **Conservation Score**
  - Evolutionary conservation of residues in proteins obtained from homology

- **Mutation Analysis**
  - Predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids.

•Inner-core residues more conserved than those at periphery

• Mutation of inner core residues are more likely to be deleterious

•Suggests possible critical functional role of those residues – e.g., as ligand binding sites or for imparting structural stability

•Relevant for pharmaceutical treatment of infectious diseases:

Core-analysis may help in identifying target sites in pathogen proteins for devising ligands to bind to those sites