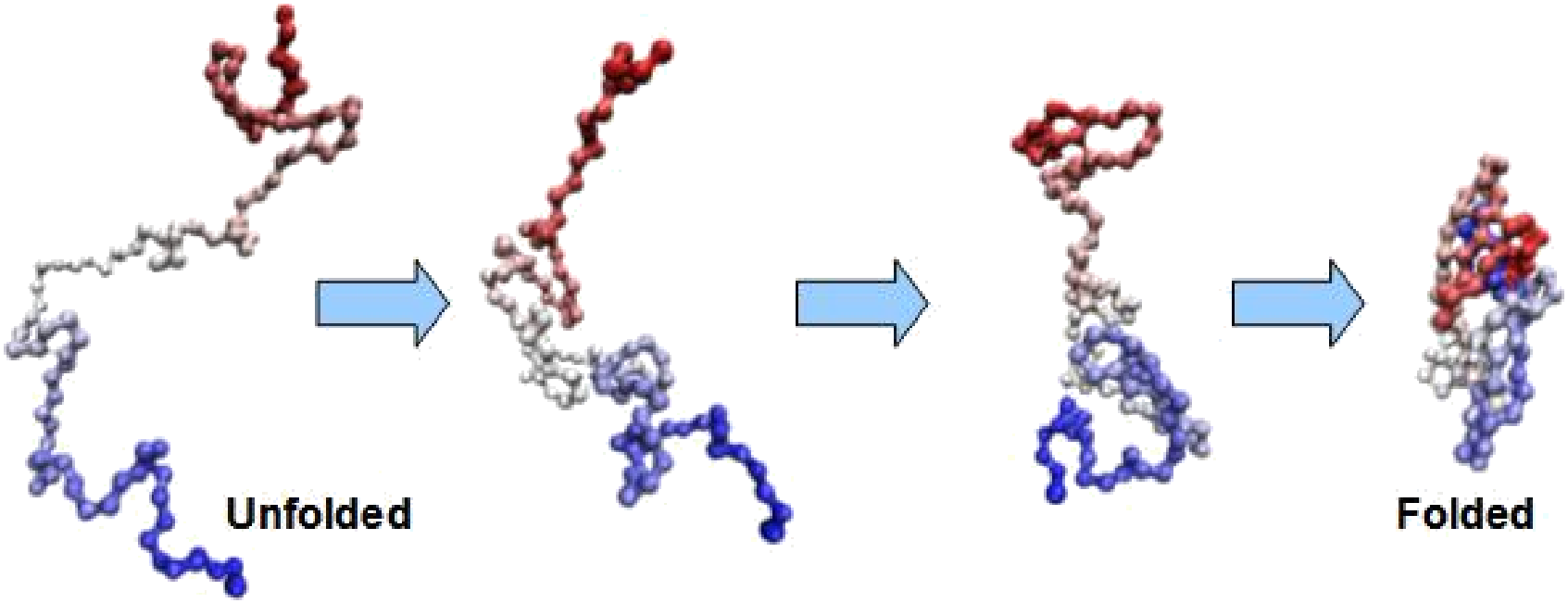
Systems Biology: A Personal View
VII. Proteins as Networks

Sitabhra Sinha

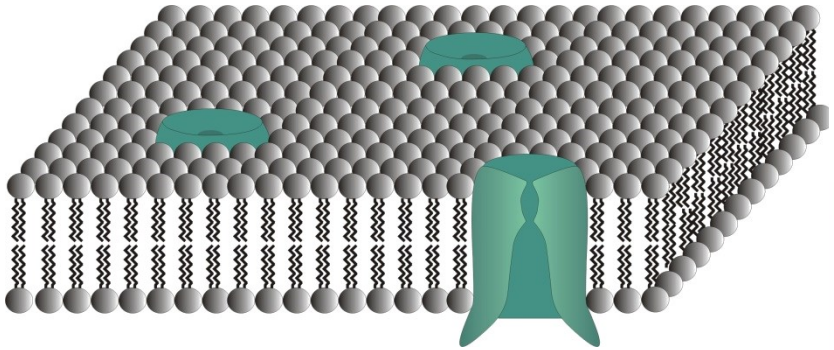
IMSc Chennai

Molecular Networks

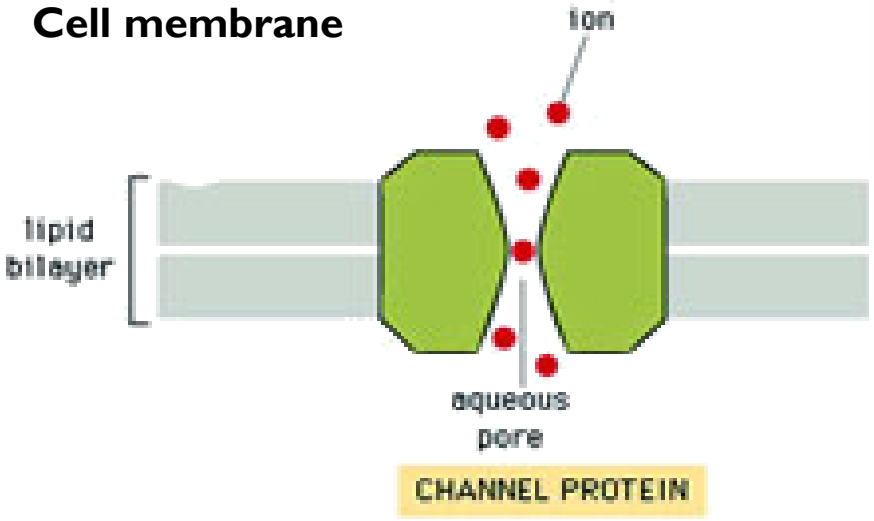


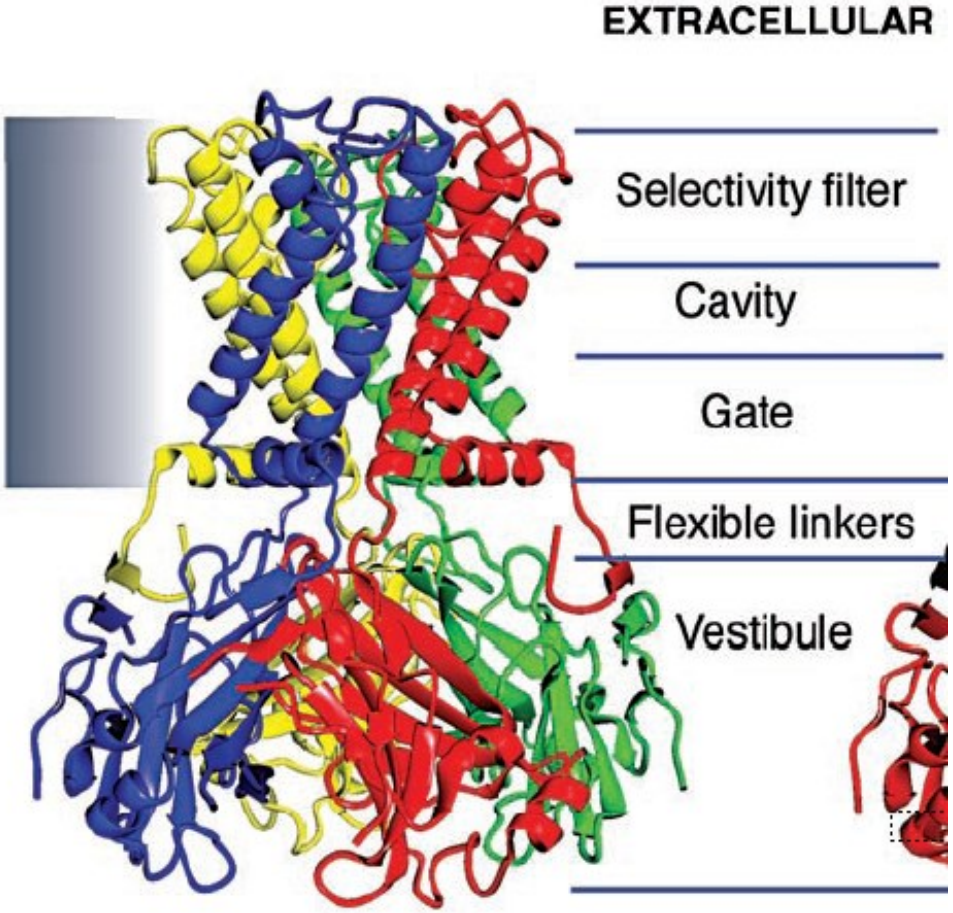
Protein Structure ≡ Network of non-covalent interactions (*links*) between amino acids (*nodes*)

Example: Kirbac1.1 Potassium ion channel protein



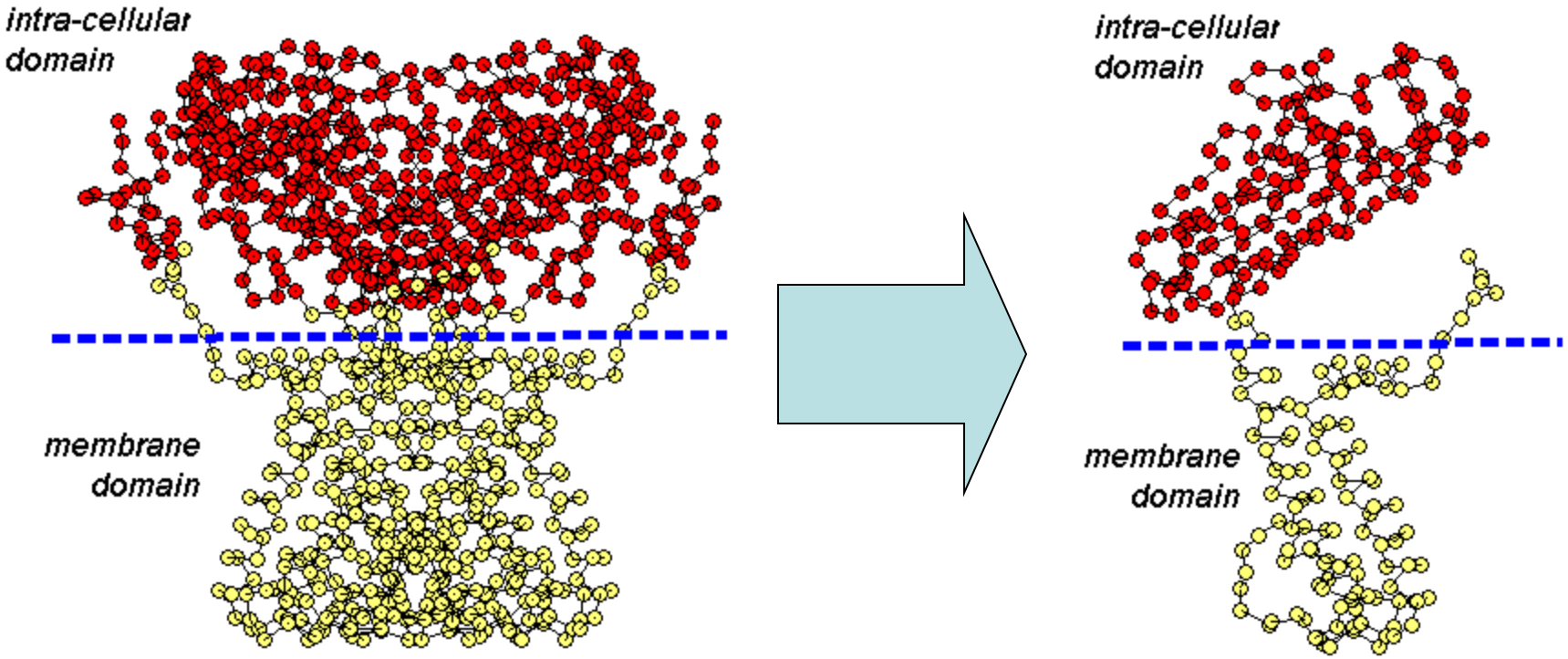
Cell membrane





Kuo et al, Science 2003

Comprises 4 identical sub-units



To construct the protein contact network from the structural data…

…obtain the x,y,z coordinates from the PDB data…

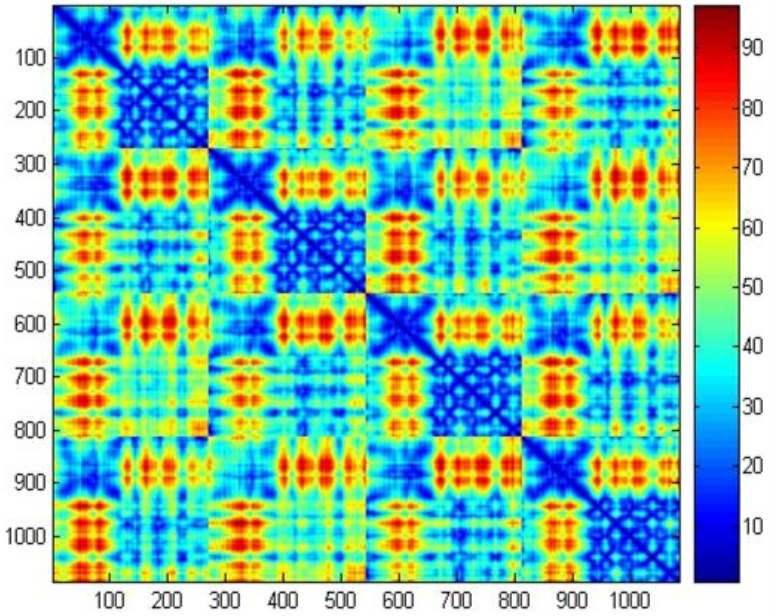
```
ATOM      1  CA  ALA A   1      28.763  10.248   6.601  1.00138.36
ATOM      2  CA  ALA A   2      30.199   7.959   3.881  1.00137.91
ATOM      3  CA  TYR A   3      30.154   4.251   2.899  1.00136.35
ATOM      4  CA  GLY A   4      31.884   1.117   1.530  1.00132.72
ATOM      5  CA  MET A   5      29.457  -1.814   0.761  1.00128.15
ATOM      6  CA  PRO A   6      27.963  -3.905  -2.144  1.00124.35
ATOM      7  CA  ALA A   7      26.076  -2.321  -5.013  1.00116.33
ATOM      8  CA  SER A   8      25.197  -2.849  -8.667  1.00108.62
ATOM      9  CA  VAL A   9      24.811  -0.380 -11.507  1.00102.08
ATOM     10  CA  TRP A  10      21.424  -1.677 -12.485  1.00 95.64
ATOM     11  CA  ARG A  11      19.412  -0.786  -9.314  1.00 89.76
ATOM     12  CA  ASP A  12      21.387   2.395  -8.871  1.00 84.17
ATOM     13  CA  LEU A  13      19.765   3.199 -12.185  1.00 77.84
ATOM     14  CA  TYR A  14      16.119   2.694 -11.149  1.00 74.00
ATOM     15  CA  TYR A  15      17.090   5.111  -8.432  1.00 77.83
ATOM     16  CA  TRP A  16      17.712   7.873 -10.908  1.00 81.88
ATOM     17  CA  ALA A  17      14.716   6.754 -12.852  1.00 77.37
ATOM     18  CA  LEU A  18      12.502   7.622  -9.913  1.00 75.09
ATOM     19  CA  LYS A  19      14.470  10.565  -8.538  1.00 77.44
ATOM     20  CA  VAL A  20      15.112  12.668 -11.615  1.00 74.87
ATOM     21  CA  SER A  21      13.044  15.559 -12.856  1.00 77.87
ATOM     22  CA  TRP A  22      10.306  14.848 -15.319  1.00 79.08
ATOM     23  CA  PRO A  23      12.028  16.445 -18.252  1.00 72.73
ATOM     24  CA  VAL A  24      15.387  15.052 -17.476  1.00 68.77
ATOM     25  CA  PHE A  25      13.321  11.945 -17.353  1.00 69.14
ATOM     26  CA  PHE A  26      11.769  11.914 -20.808  1.00 71.69
ATOM     27  CA  ALA A  27      15.000  13.351 -22.179  1.00 72.26
ATOM     28  CA  SER A  28      16.657  10.279 -20.708  1.00 74.27
ATOM     29  CA  LEU A  29      14.272   8.082 -22.671  1.00 73.90
ATOM     30  CA  ALA A  30      14.534  10.234 -25.774  1.00 73.72
ATOM     31  CA  ALA A  31      18.340   9.977 -25.691  1.00 73.22
ATOM     32  CA  LEU A  32      17.834   6.233 -25.246  1.00 71.91
ATOM     33  CA  PHE A  33      15.278   6.123 -28.077  1.00 74.33
ATOM     34  CA  VAL A  34      17.902   7.573 -30.345  1.00 78.38
ATOM     35  CA  VAL A  35      20.664   5.290 -29.183  1.00 82.23
ATOM     36  CA  ASN A  36      18.229   2.387 -29.468  1.00 91.83
ATOM     37  CA  ASN A  37      17.158   3.403 -32.983  1.00 98.81
ATOM     38  CA  THR A  38      20.717   3.590 -34.260  1.00 99.51
```

… and calculate the (Euclidean) distance between each pair of amino acids…

For any pair $P = (p_x, p_y, p_z)$ and $Q = (q_x, q_y, q_z)$

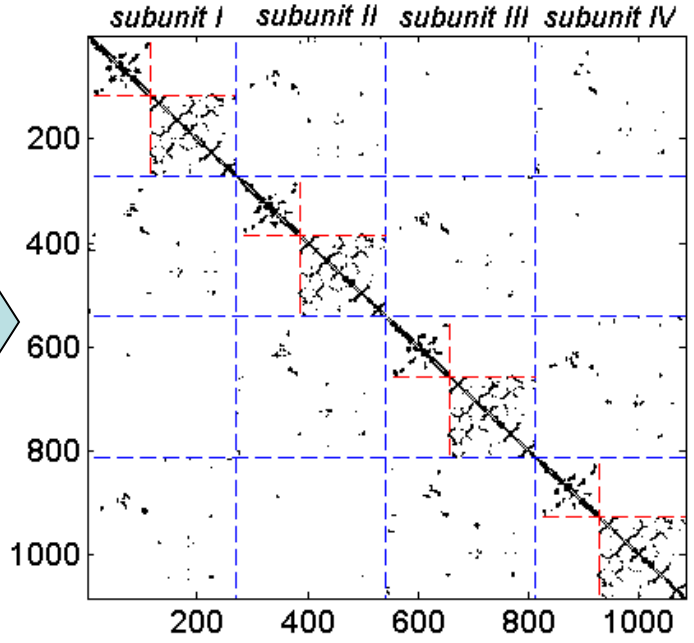
the distance is calculated as: $\sqrt{(p_x - q_x)^2 + (p_y - q_y)^2 + (p_z - q_z)^2}$.

…to obtain the Distance matrix …

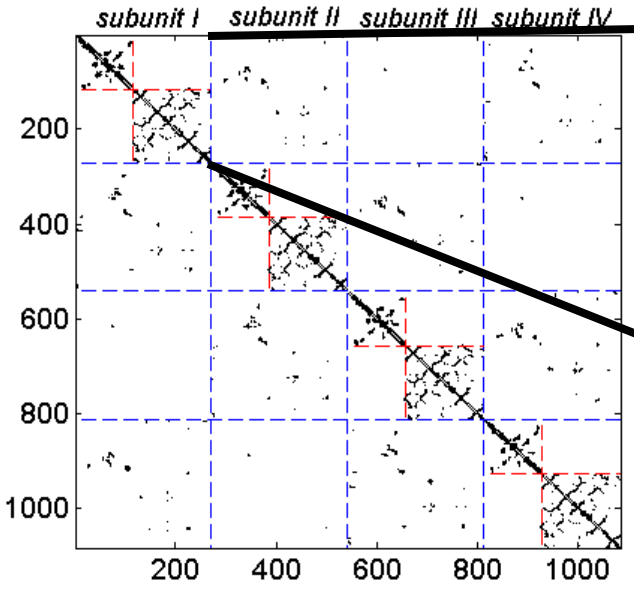


Threshold
Cutoff = 12 A

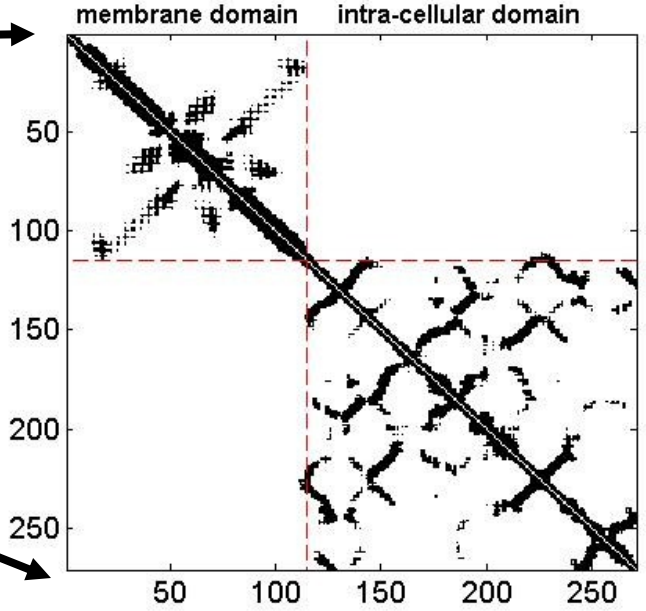
and the Adjacency matrix

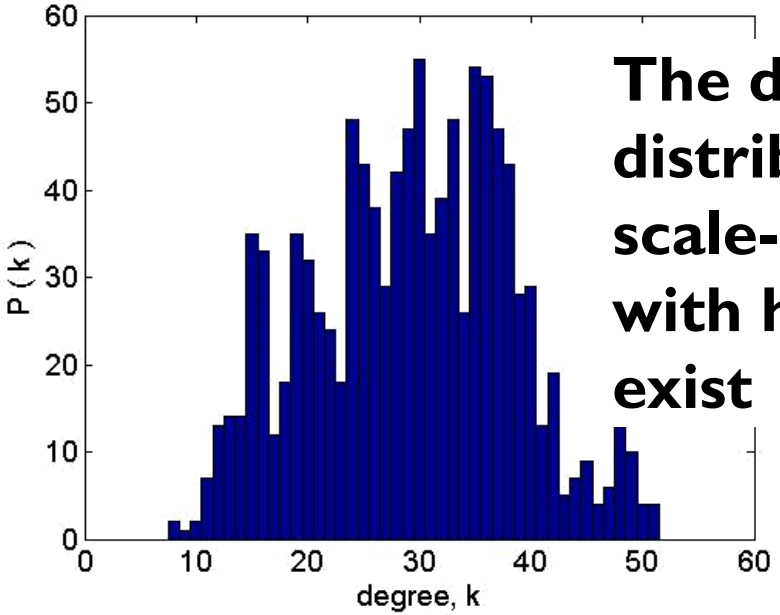


Protein Contact Network

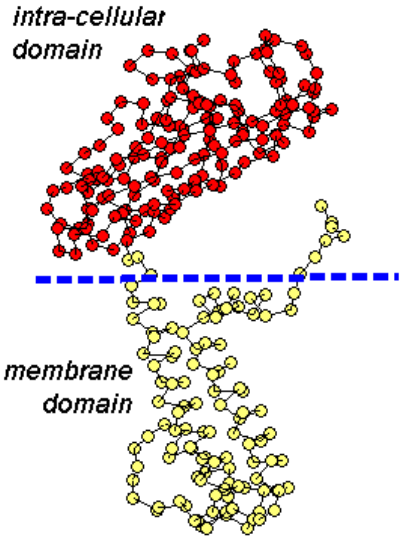


Magnification of a sub-unit reveals modular structure



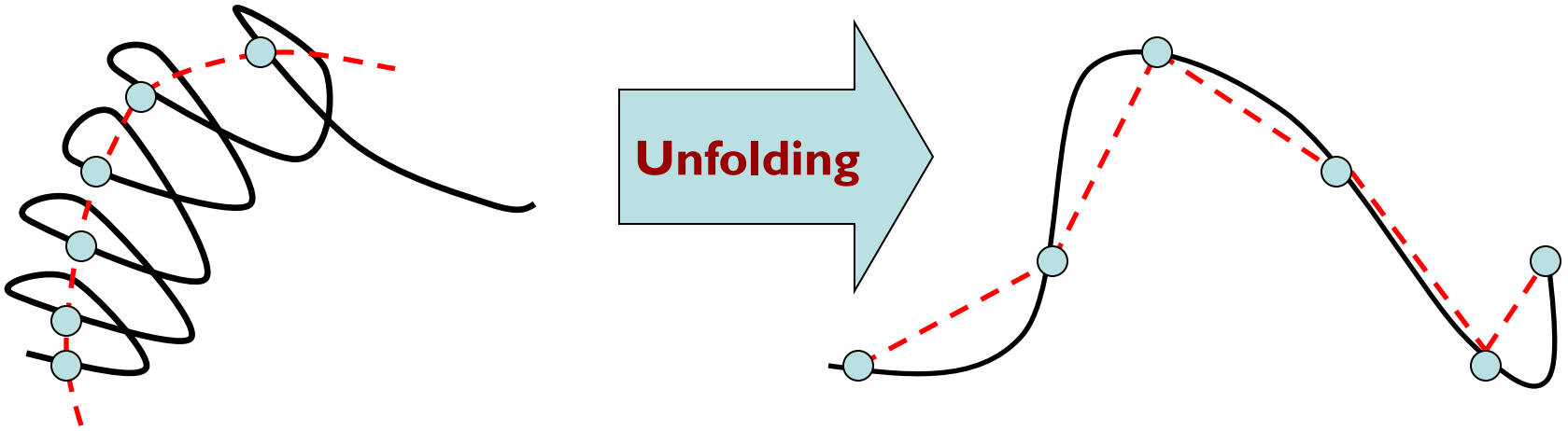


The degree distribution is not scale-free but nodes with high degree do exist



Is the protein contact network small-world ?

Yes, low average path length and high clustering



The genesis of small-world nature is from the existence of cross-links as a result of the folding of the protein

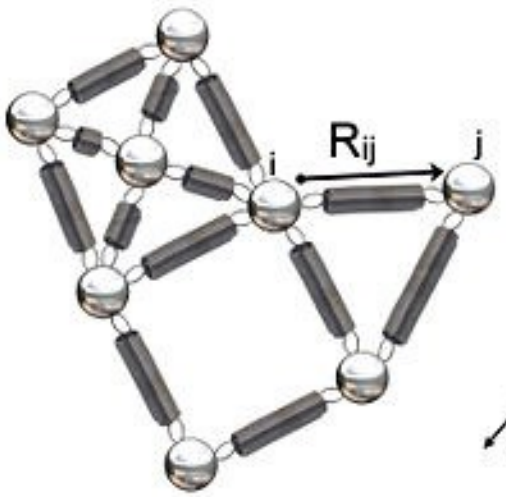
Is the small-world nature of a protein functionally important ?

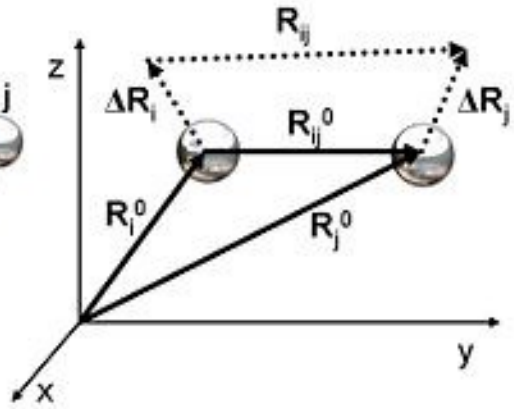
The cross-links provide structural stability

Understanding Protein dynamics from network analysis

Protein = elastic network of balls (C-α atoms) connected by springs (chemical interactions)

Source: Wikipedia



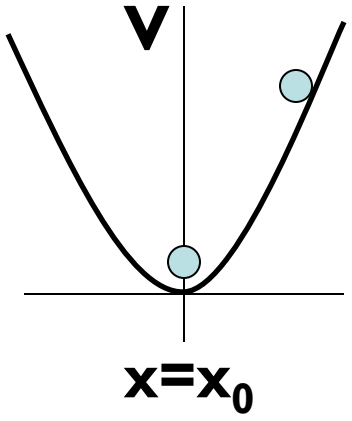


Under the

Harmonic potential approxn:

$V(x) \approx V(x=x_0)+(1/2)(x-x_0)^2\ \partial^2V/\partial x^2 + \ldots$

[Force = $\partial V/\partial x = 0$ at $x = x_0$]



PE of network, $V = (k/2)\ \Sigma_{i,j=1\ldots N}\ (R_{ij} - R^0_{ij})^2$

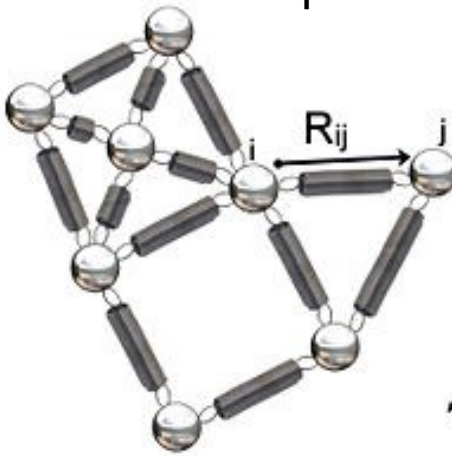
$V = (k/2)\ \Sigma_{i,j=1\ldots N}\ (\Delta R_i - \Delta R_j)2$,

where $R_{ij} = R_i - R_j = (R^0_{ij} + \Delta R_i - \Delta R_j)$

k: force constant

Understanding Protein dynamics from network analysis

Source: Wikipedia



Under the
Harmonic potential approxn:

PE of network, $V = (k/2)\ \Sigma_{i,j=1\ldots N}\ (\Delta R_i - \Delta R_j)^2$,

where $R_{ij} = R_i - R_j = (R^0_{ij} + \Delta R_i - \Delta R_j)$

Or,

PE of network, $V = (k/2)\ \mathbf{(dR)^T\ L\ (dR)}$

dR: column vector of fluctuations, i.e., displacements from eqlbm

L: *Laplacian* or *Kirchoff* matrix
off-diagonal elements L(i,j) = -1, if d(i,j) < cut-off; L(i,j) = 0, otherwise
diagonal elements L(i,i) = degree k (i) = sum of all links for node i

Correlations between fluctuations,

$< dR(i).dR(j) > = (k_B T/k)*L^{-1}(i,j)$

The vibrational normal modes of the protein are governed by the eigenvalues of **L** : small eigenvalue implying large-scale motion

The Graph Laplacian

See M E J Newman, *Networks*, Section 6.13

Consider diffusion processes on networks – i.e., a process by which something (a contagion, a signal or an idea) spreads across a network.

Let this "something" exist initially in varying quantities (say randomly chosen) on the different noes of a network, with the amount in node i being denoted X_i.

Also let this "something" **diffuse** along the links, flowing from node j to an adjacent node i at a rate governed by the "density gradient" $C(X_j - X_i)$ where C is the *diffusion constant*.

$\Rightarrow$ the rate at which X_i is changing is $dX_i/dt = C \Sigma_j A_{ij} (X_j - X_i)$

$\Rightarrow dX_i/dt = C \Sigma_j A_{ij} X_j - C X_i \Sigma_j A_{ij} = C \Sigma_j A_{ij} X_j - C X_i k_i = C \Sigma_j (A_{ij} - \delta_{ij} k_i) X_j$

Thus, in matrix form $d\boldsymbol{X}/dt = C (\boldsymbol{A} - \boldsymbol{D}) \boldsymbol{X} = - C \boldsymbol{L} \boldsymbol{X}$

where

$\boldsymbol{A}$: Adjaceny matrix, $\boldsymbol{D}$: diagonal degree matrix, and, $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$ is the Laplacian matrix

The diffusion equation can be solved in terms of the eigenvectors $\boldsymbol{v}_i$ of the Laplacian $\boldsymbol{L}$:

$\boldsymbol{X}(t) = \Sigma_i a_i(t) v_i$ where the time evolution of the coefficients a_i can be expressed in terms of the eigenvalues $\lambda = \{\lambda_i\}$ of the Laplacian $\Rightarrow a_i(t) = a_i(0) \exp(-C \lambda_i t)$

All eigenvalues of the Laplacian matrix are non-zero, the smallest being $\lambda_1 = 0$

corresponding to the eigenvector $\mathbf{1} = \{1, 1, 1, 1, \ldots, 1\}$

Gaussian Network Model of Protein dynamics

See: Wikipedia entry

Tirion (1996)

Potential energy of the network (under harmonic approximation):

$$V_{GNM} = \frac{\gamma}{2}\left[\sum_{i,j}^{N}(\Delta R_j - \Delta R_i)^2\right] = \frac{\gamma}{2}\left[\sum_{i,j}^{N}\Delta R_i \Gamma_{ij} \Delta R_j\right] = \frac{\gamma}{2}[\Delta X^T \Gamma \Delta X + \Delta Y^T \Gamma \Delta Y + \Delta Z^T \Gamma \Delta Z]$$

Assuming that : Probability distribution of fluctuations is Gaussian…

$$p(\Delta X) \propto exp\left\{-\frac{\gamma}{2k_B T}\Delta X^T \Gamma \Delta X\right\} = exp\left\{-\frac{1}{2}\left(\Delta X^T \left(\frac{k_B T}{\gamma}\Gamma^{-1}\right)^{-1} \Delta X\right)\right\}$$

Including normalization constant $p(\Delta X) = \frac{1}{\sqrt{(2\pi)^N \frac{k_B T}{\gamma}|\Gamma^{-1}|}} exp\left\{-\frac{1}{2}\left(\Delta X^T \left(\frac{k_B T}{\gamma}\Gamma^{-1}\right)^{-1} \Delta X\right)\right\}$

… and isotropic

$$P(\Delta R) = p(\Delta X)p(\Delta Y)p(\Delta Z) = \frac{1}{\sqrt{(2\pi)^{3N}|\frac{k_B T}{\gamma}\Gamma^{-1}|^3}} exp\left\{-\frac{3}{2}\left(\Delta X^T \left(\frac{k_B T}{\gamma}\Gamma^{-1}\right)^{-1} \Delta X\right)\right\}$$

Therefore correlation between fluctuations can be evaluated from the covariance

$$< \Delta X \cdot \Delta X^T >= \int \Delta X \cdot \Delta X^T p(\Delta X) d\Delta X = \frac{k_B T}{\gamma}\Gamma^{-1} =< \Delta Y \cdot \Delta Y^T >=< \Delta Z \cdot \Delta Z^T >= \frac{1}{3} < \Delta R \cdot \Delta R^T >$$

Correlations between fluctuations, $< dR(i).dR(j) > = (k_B T/\gamma)*\Gamma^{-1}(i,j)$

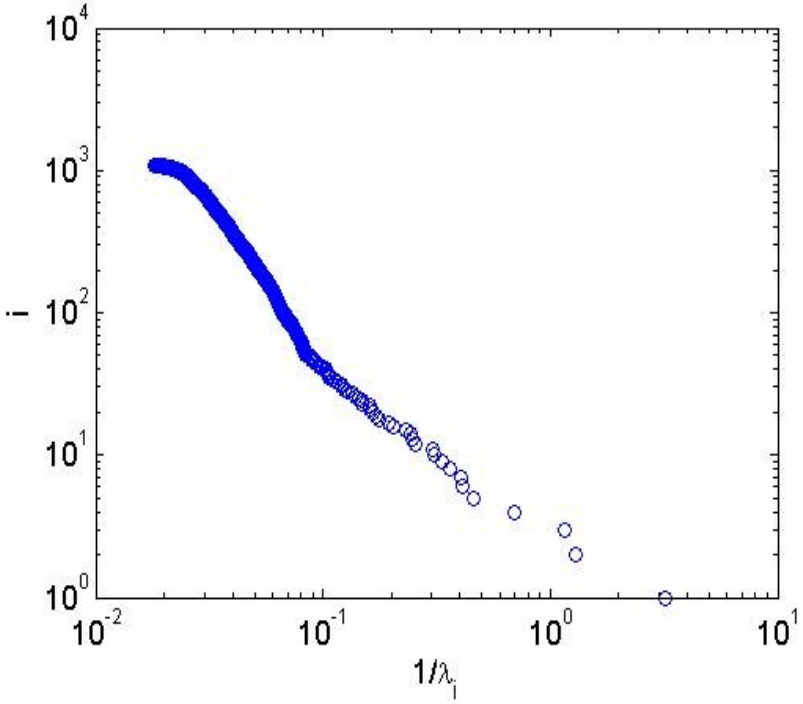
The vibrational normal modes of the protein are governed by the eigenvalues of Γ: small eigenvalue implying large-scale motion

Understanding Protein dynamics from network analysis

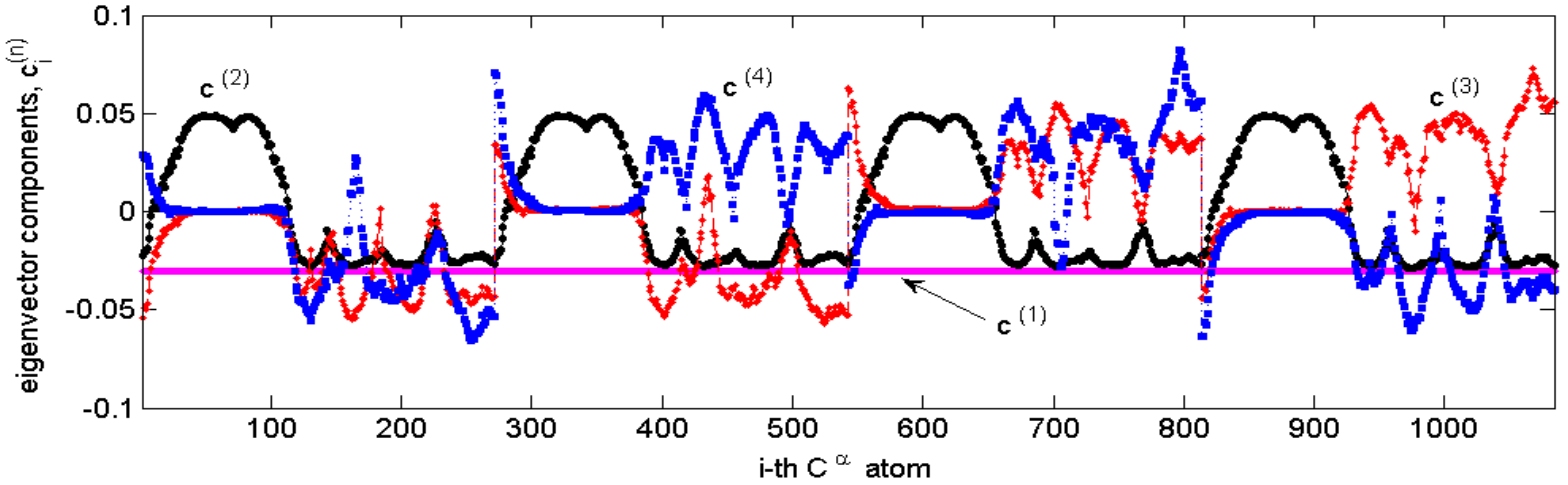
The spectrum of eigenvalues of L for Kirbac1.1 protein

4 very small eigenvalues indicate dominance of largest scale motion by 4 sub-units.

Other large scale motions: possibly dominated by modular structure



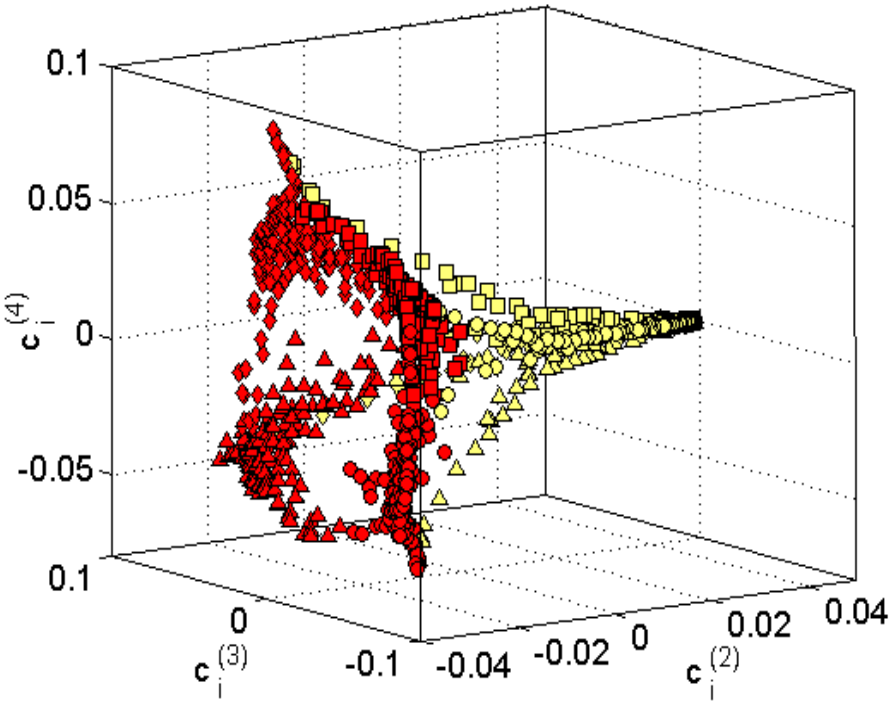
The eigenvector components of the smallest eigenvalues of **L**



Understanding Protein dynamics from network analysis

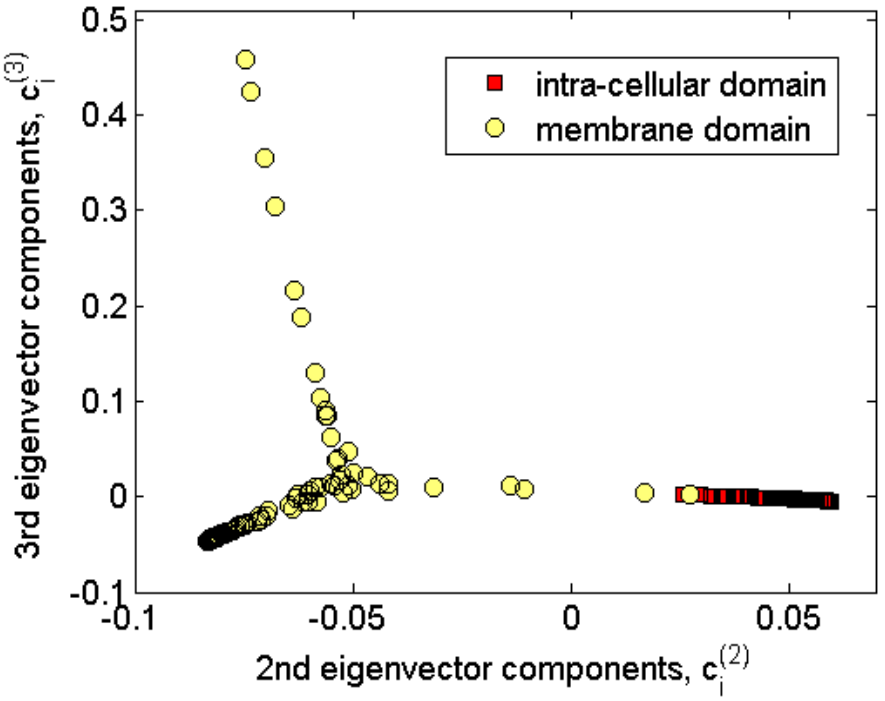
The eigenvector components corresponding to the smallest non-zero eigenvalues indicate how the module motions are correlated

Similar analysis of the Internet in Eriksen et al, PRL 90 (2003) 148701



One of the four sub-units

The entire protein



But the Protein Contact Network also contains links that correspond to the backbone…

Long-range Interaction Network (LIN)

…which does not give us much information about the folded tertiary structure of the protein

To focus on the cross-links, we need to construct the

obtained from PCN by excluding links among spatially neighboring nodes along the backbone

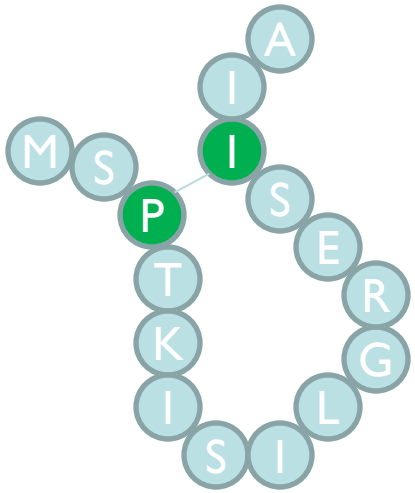
Example:

LIN may be constructed from PCN by removing links between nodes corresponding to a *cumulative spatial distance* ≤ 10Å.

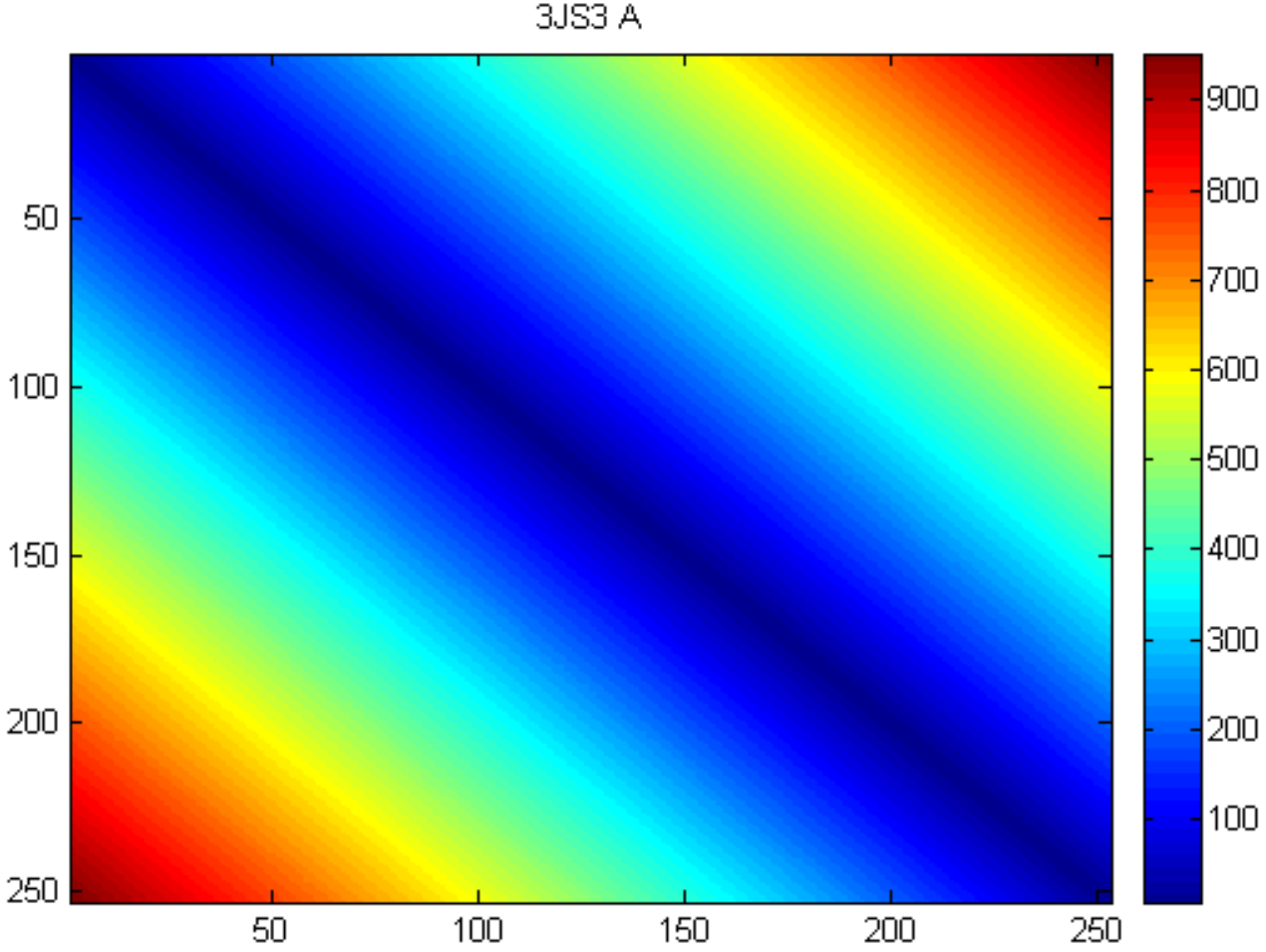
First, we obtain the

Cumulative Distance Matrix (CDM)

i.e., Euclidean distances between all pairs of C-α atoms



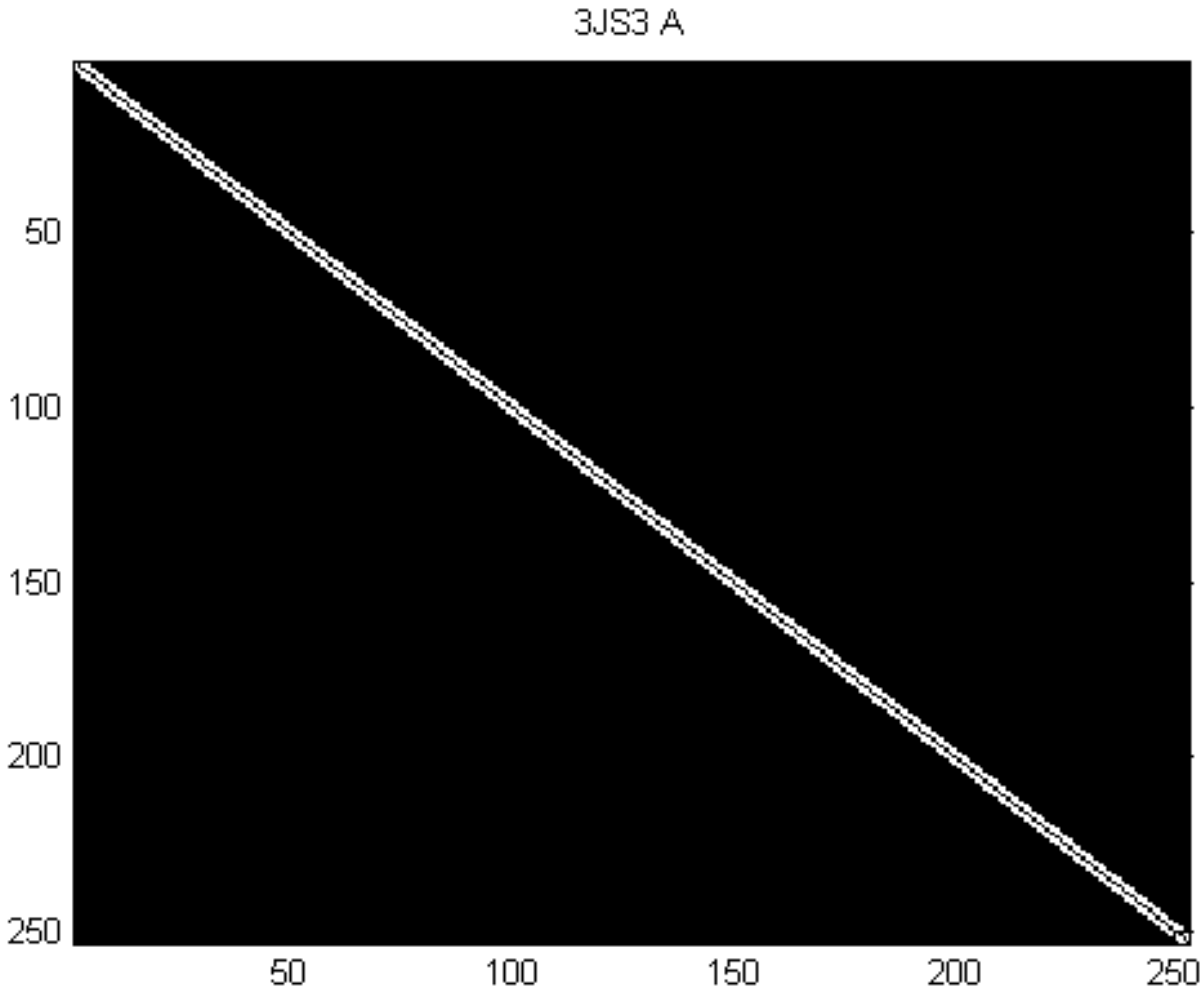
Cumulative distance from M to P = distance from M to S + distance from S to P



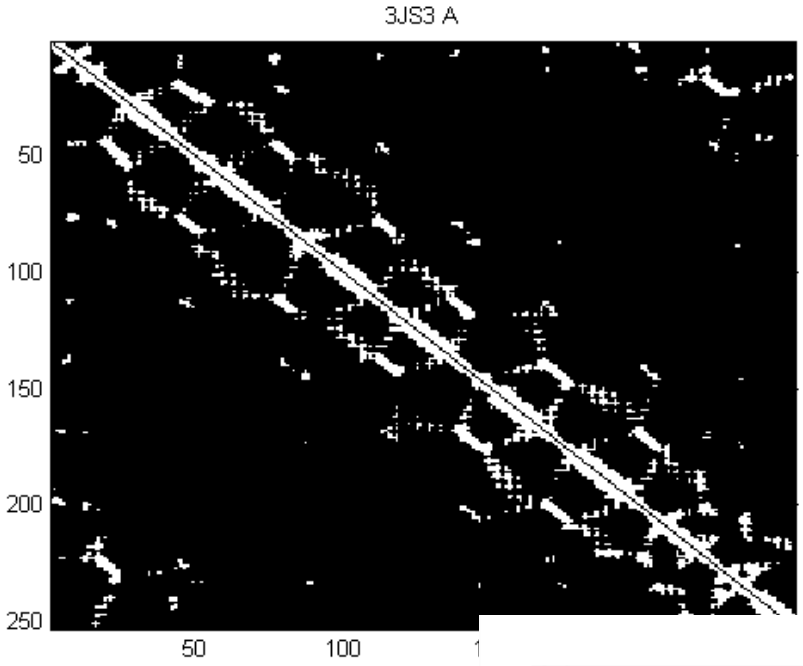
Next, we obtain the

Backbone Adjacency Matrix (BAM)

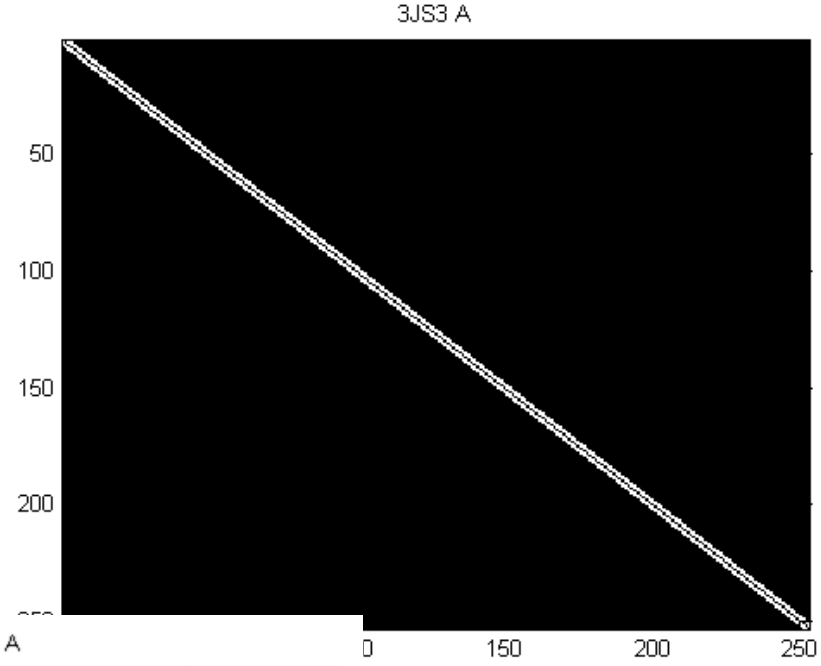
from the CDM by retaining only those links corresponding to Euclidean distance < 10 A



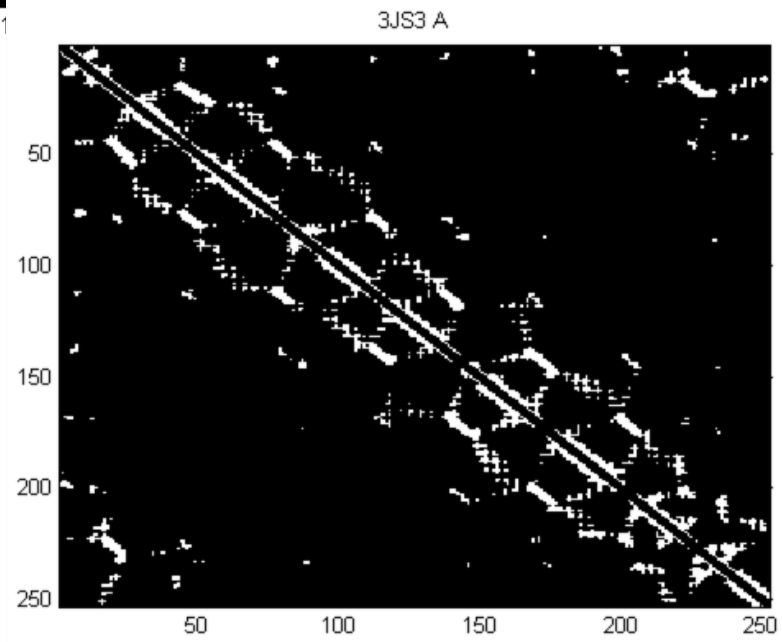
Finally, the Long-range Interaction Network (LIN) is obtained by keeping those links in PCN which do not appear in BAM



PCN



BAM



LIN