## Lecture 7: Protein Collective Motions

**I. Introduction**

The focus of this lecture (and lab) is large conformational changes in proteins. Previously we have been concerned with small conformational changes in proteins that occur on the picosecond to nanosecond time scale and that involve distance amplitudes of less than 10 Å. These types of motions are described as exposed side chain rotations and loop motions in the table below. Now we turn our attention to slower conformational changes involving larger distance amplitudes.

| Time scale | | Amplitude | Type |
|---|---|---|---|
| femto - picoseconds | $10^{-15}$ s - $10^{-12}$s | 0.001 Å - 0.1 Å | Bond stretching<br>Angle bending |
| pico - nanoseconds | $10^{-12}$ s - $10^{-9}$s | 0.1 Å - 10 Å | Exposed side chain rotation, loop motion<br>MD |
| nano - microseconds | $10^{-9}$ s - $10^{-6}$s | 1 Å - 100 Å | Helix - coil transition, hinge bending<br>Collective Motions |
| micro - seconds | $10^{-6}$ s - $10^{-0}$ | 10 Å - 100 Å | Macroscopic fluctuations, protein folding |

Although many protein enzymes and receptors carry out their respective functions by a shift in the conformational ensemble involving slight conformational changes, many other proteins experience large conformational changes during their respective functions. The study of protein motions has increased exponentially in recent years. It is now recognized that to understand the function of many proteins we must understand their dynamic fluctuations at multiple time scales. Only recently have traditional simulation techniques provided trajectories that correspond to μs and longer. These conventional MD methods are feasible only for small proteins given today's computational limitations. So, we use other methods to study large conformational changes.

Large conformational protein motions, by definition, involve **collective motion** of multiple residues simultaneously and this is the next topic of interest.

**II. Normal Modes.** *Atoms or groups that oscillate with the same frequency make a collective normal mode*. The very low frequency modes describe the large amplitude motions of the protein and these are the ones we are interested in.

**The low frequency normal modes are believed to describe motions that may be related to the function of a protein.** The collective motion of atoms with similar frequencies can be displayed to visualize these modes.

An example showing the four lowest frequency normal modes of a polypeptide helix will be shown in class. In the first two lowest frequency modes the helix bends in two different directions, in the next lowest frequency mode it rotates about a single internal joint and in the next mode it rotates about two internal joints. Again, in each of these modes different groups of atoms move with the same frequency.

The goal of protein normal mode analysis is to predict motions which *may* be involved in the functioning of the protein. Research has shown that a combination of the lowest frequency modes in many cases will predict actual protein motions captured by X-ray crystallography.

**III. Methods to study large protein motions**.

**A. X-ray Crystallography**. A number of X-ray crystal structures of proteins show that a single protein can crystallize in more than one conformation. These structures give us a time averaged view of separate energy regions of the ensemble of structures. Frequently the different structures are due to the presence or absence of a ligand or other complexing protein.

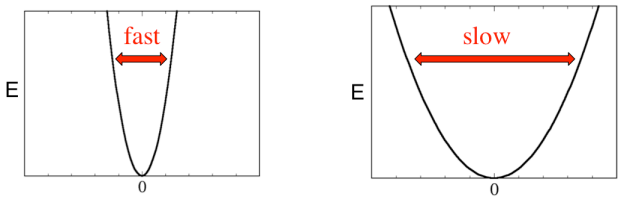
**B. Normal Mode Analysis.** What if only one structure for a protein is available? Is it possible to predict large conformational changes that may be involved in the function of that protein? This problem is especially important in the era of structural genomics when we may be presented with many protein structures for which the function is not known. Sometimes knowing the conformational changes a protein experiences will help to identify the binding site and therefore its function. In the genomics era we need a method that is much faster than traditional MD simulations for " high throughput" analysis.

One method to help one predict large conformational change from a single structure is normal mode analysis. Normal mode analysis is a method to model the **collective motions** of atoms or groups in a molecule. It treats atoms or groups of atoms as harmonic oscillators coupled together by springs. It is assumed that the pairwise relative atomic motion is described by a harmonic potential energy function around a local energy minimum. You will remember that we modeled bond stretching and bond angle deformation previously by using harmonic potentials. The next figure illustrates that this Hookian-type of potential function can be related to the frequencies of bond and angle vibrations as well as the potential energies of deformations.

$E = K((b - b_0)^2$
K = Spring constant
K ~ 500 kcal/mol/Å$^2$

$E = K(\theta - \theta_0)^2$
K ~ 50 kcal/mol/Å$^2$




Another way to say the above is that the spring constant, K, is related to the frequency of vibrations. Atoms that vibrate with the same frequencies can be modeled with the same spring constants. The goal of a normal mode analysis is to calculate the spring constants (and therefore the vibrational frequencies) of each atom or group of atoms in a protein.

A molecule made up of N atoms will have 3N-6 normal modes differing in frequency and direction. Six modes of motion are described by whole molecule rotation and translation and we exclude those. We are interested in the *relative* motion of atoms to other atoms in the molecule. Different groups of atoms may oscillate with the same frequency. Each such collective oscillation is a **normal mode**. To calculate normal modes, we want the direction and amplitude of oscillation of atoms for each frequency. Atoms oscillating at the same *frequency* are then grouped together to describe a mode.

Note the thinking process here. We are interested in atoms that oscillate with the same frequency, not groups of atoms that *cause* each other to move together. We are looking for a statistical correlation. Even if correlated oscillations do not have a causal relationship, if they occur together in a protein there will be intervening residues also with correlated motion. These intervening residues describe the conformational linkage between correlated residues. Interrupting this linkage in the protein may destroy the global network of interactions and destroy the correlation. These relationships are important in protein design.

One can conceptualize the method of normal mode simulation by the following steps:

1. Randomly displace atoms small distance from energy *minimized* structure.
2. Use standard force field to calculate energies between atoms.
   $E = \varepsilon_{ij}[(R_{min} / R_{ij})^{12} - 2(R_{min} / R_{ij})^6] + q_iq_j / 4\pi Dr_{ij}$
3. Use energies to calculate effective harmonic spring constants and frequencies.
   $E = K(R_{ij} - R_{min})^2$, where K $\propto$ frequency.
4. Repeat 1, 2, 3 many times.
5. Cluster atoms with same frequencies.

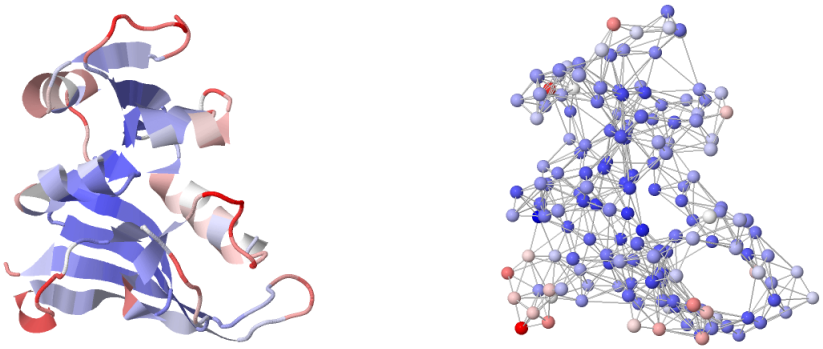
Again, the goal is to identify atoms that move with the same frequency; that is the definition of a normal mode. The steps above are simplified to allow one to conceptualize the interplay between energy potentials as we have previously described them in class. In fact, the actual calculation of normal modes uses *forces* derived from the interatomic energies to get out the frequencies.

The method just described has historical importance but, in practice, is not used today. I have included it because I think it helps to understand this topic. It turns out to be quite time consuming and it was discovered that an alternate method to estimate normal modes is much faster. This second method is described next.

**C. Gaussian Network Model**. This analysis, sometimes referred to as the **Gaussian** or **elastic network analysis**, takes into account the number of nearest neighbor residues for each residue in a protein. Those residues which have a large number of nearby residues will have a larger local packing density and it would be more energetically unfavorable to displace these residues. **In fact, the shape of the normal mode energy potential for residue fluctuations should be related to the local packing density and geometry of a residue**. This relationship allows one to estimate the normal modes of each residue in a single protein structure.

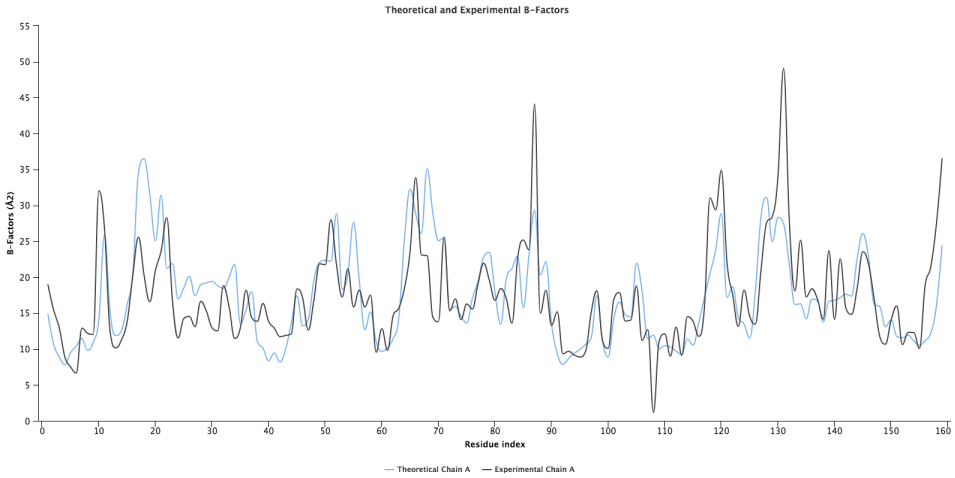
One way to visualize how this analysis is done is shown in the following figure. On the left is a ribbon diagram of a protein. On the right is a network of lines between each residue Cα and its nearest neighbors in the same protein. Imagine that each line is a Hookean spring with spring constant associated with stretching. The energy required to displace each residue is related to the number and length of network links attached to each residue.



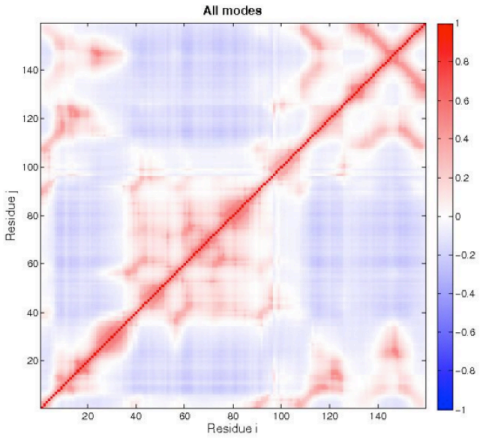
From the energy required to displace each residue we can calculate which residues should have similar energies, and therefore similar Hookean spring constants, and therefore frequencies, of fluctuations. Such groups should be correlated in their motions and be members of the same normal mode.

A validation of the Gaussian network analysis method is that the calculated apparent B-factors agree very well with the experimental B-factors as determined by X-ray crystallography. Below is a plot of 1RA2 B-factors obtained from the two methods.

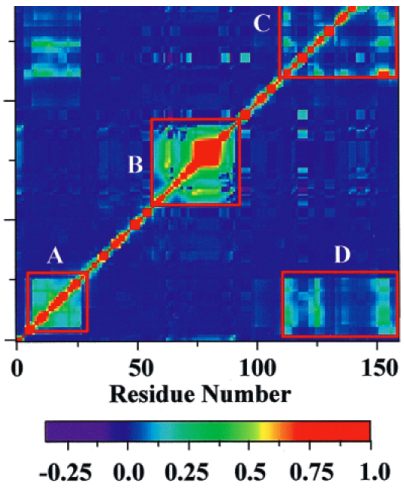
 




One can also look at the motion correlations of residues in structure 1RA2 as shown in the following figure on the left. As you can see from a comparison to the plot on the right the residue *motion* correlations (left) are somewhat consistent with the residue *stability* correlations (right) as calculated using a free energy force field applied to an ensemble of partially unfolded dihydroreductase protein.




http://ignm.ccbb.pitt.edu/




**D. Network Analysis of MD trajectories.** In this analysis each residue is considered a node in a network with edges represented by lines between the nodes. The network model looks somewhat like the elastic network model above. The difference is that edges in the elastic network model are drawn between all neighbor residues; in the network calculated from the trajectory, the edges represent only those residues that are correlated in their motions during the trajectory. *Here the correlated residue may be distant from each other in Cartesian space unlike in the Gaussian network model above.*

We can also apply a value to each edge that is proportional to the correlation between each pair of nodes for visual inspection. From rigorous analysis of the edge values, clusters of residues that move together are identified and called communities. Finally, the nodes between communities that greatly affect properties of the network represent residues thought to be important for allosteric communication between communities and

from these the allosteric linkage in the protein can be identified.

To calculate the correlation of residues during a molecular dynamics trajectory we apply the following steps:
1. Calculate the standard deviation of positional displacement in each direction (here for the *x* direction where the bar-*x* is the mean),

$$SD_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

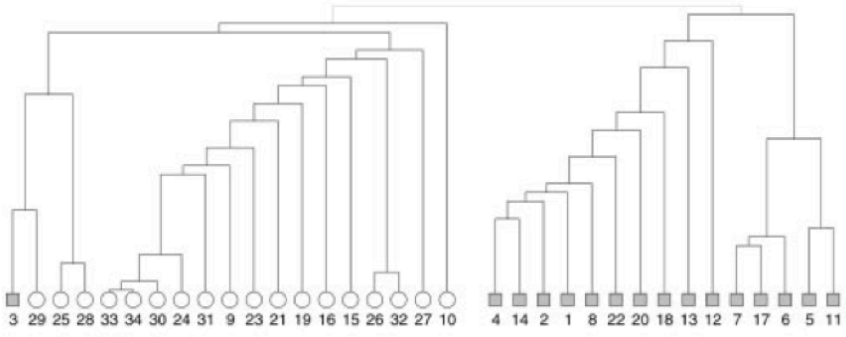
2. Then the covariance between two directions of motion is calculated (here for the *x*, *y* directions)

$$Covariance_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

3. And finally the correlation is the covariance divided by the product of the two standard deviations,

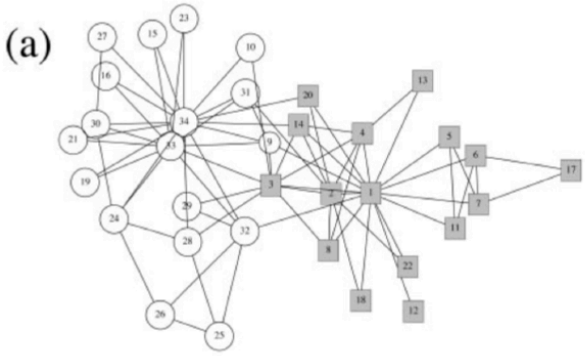
$$Correlation = \frac{Covariance}{SD_x \times SD_y}$$

One then uses a clustering-like algorithm to identify groups of residues with similar correlations. You may be familiar with the standard hierarchical clustering algorithm where one starts by identifying the two data items that are most similar (items 33 and 34 in the data set shown below). These two items are grouped to form a cluster of two. Then the next most similar data item is added to that to form a cluster of three with a correlation equal to the average of the cluster members, etc. This process is continued until all data items are included. The lengths of the vertical lines in the figure indicate the degree of relatedness between the data items or clusters. Note that item (3) is an outlier.

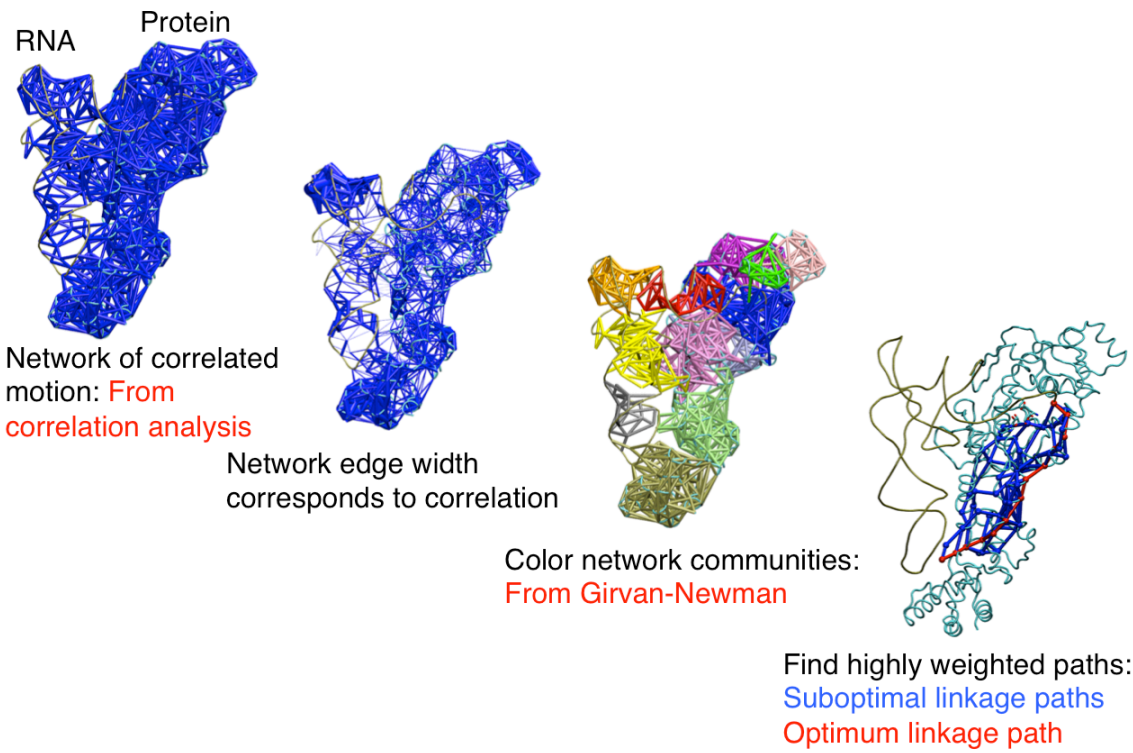



Network communities are identified by a different algorithm called the Girvan and Newman method. In this method a network of all data items connected by all other data items is made initially. Then the edges with highest "**betweenness**" are removed. This is a hard concept to understand but the idea is that we want those nodes that are highly connected to be sorted into groups that have less dense connections. The Girvan-Newman method is applicable where we want to find the shortest routes of transfer between communities. A Girvan-Newman community network of the same data as in the dendrogram above is shown in the next figure. Note that item (3), that was an outlier in

the hierarchical clustering above, is shown to be at the interface of both communities and is along a path connecting the communities. This relationship is not identified by the hierarchical clustering above.




The following figure of a tRNA/protein complex illustrates the process of network analysis.




The final image at the lower right in the above figure illustrates the allosteric linkage paths between the parts of the protein binding the RNA anticodon (lower left) and binding the CCA end of the RNA (upper right). In your simulation of a small protein there may be only one linkage path between parts of the molecule and in this case the suboptimal and optimum paths will be the same.