BITS Pilani
K K Birla Goa Campus

Project Report

PHYSICS OF PROTEIN FOLDING DYNAMICS

Submitted in partial fulfillment of Study Project (BITS F266). Semester I, 2014-15

Sajidur Rahman 2011B5A7496G
Nikhil 2011B5A8419G

Contents

Acknowledgement

We are using this opportunity to express our gratitude to Dr. P.N. Deepak, who supported us throughout the course of this project. We are thankful for his aspiring guidance, invaluably constructive criticism and friendly advice during the project work. We are sincerely grateful to him for sharing his truthful and illuminating views on a number of issues related to the project.

Thank you.

Project Description

Title

Physics of Protein Folding

Project Highlights

The importance of protein folding has been highly talked about and has been a very interesting field for research. Proteins play an inevitable role in different biological processes in a living body. By coiling and folding into a specific three-dimensional shape they are able to perform their biological function. Through this project, we aim to look at the basic physical principles that govern the folding of proteins in living cells. We try to address this problem from the point of view of Quantum Mechanics and Statistical Physics.

November 30, 2014

Project Summary

The following points sum up the work that have been done for the project:

- Protein Structure: Primary, Secondary, Tertiary, Quaternary.
- Different models discussing the folding process.
- Stochastic Processes and Random Self Avoiding Walk.
- Analogy with Quantum Chromo-dynamics (QCD) and Path-Integral Representation.
- Self-Avoiding Random Walk and Computer simulation to find compactness Index and compare it with calculated value.
- Two state kinetic model - To compare Protein folding kinetics with the standard two state kinetic model equations and finding the rate constants.
- Rouse Model for Polymer dynamics to find parameters like Angular momentum and diffusion co-efficient which are relevant to the study of Hydrophobic Collapse.

1. Protein: Introduction

Proteins are macromolecules involved in virtually all the biochemical processes which take place inside the cell, such as e.g. catalysis, charge transport, signal transduction. Clearly, understanding the physical principles which drive the internal dynamics of individual proteins and shape the protein-protein interaction would have countless implications in molecular biology, drug design and nano-biotechnology. In particular, a central open challenge at the interface of physics, biochemistry and molecular biology concerns the prediction of the protein folding pathways i.e. the sequence of conformational transitions which these molecules perform in order to reach their stable and biologically active configuration

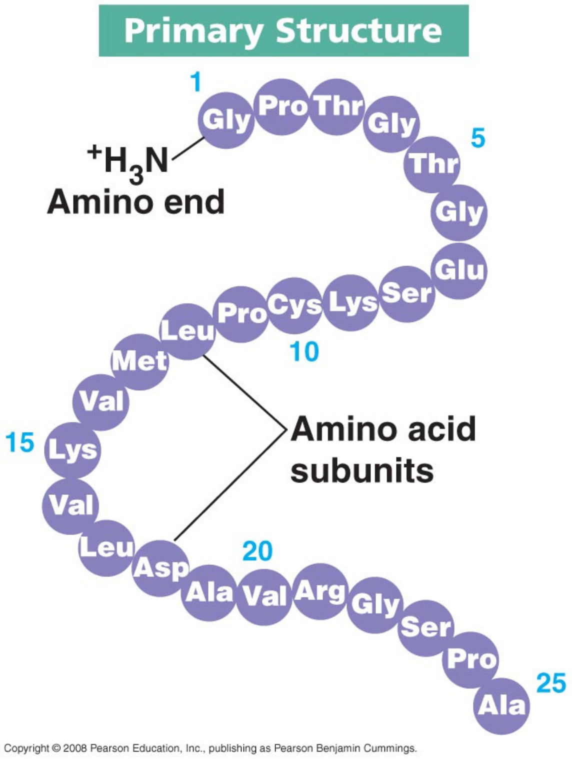
Proteins are polymers built from amino acids. There are 20 amino acids That are common to all living organisms on Earth. These are all amino acids, with the amino and carboxylic acid groups separated by a single Carbon atom. Each amino acid is distinguished by the chemistry of the side chains that are attached to the carbon atom.

2. Protein: Structure

The fibrous structural proteins contain very long helices. Water-soluble and membrane proteins, however, are more globular in shape and therefore must adopt more complex folds. The overall global conformation of a protein is its tertiary structure. Although tertiary structures are highly varied across different proteins, there remains some degree of regularity within them, because local regions adopt regular secondary structures. Groups of helices associate to form super secondary structures, which are motifs that recur frequently in many different proteins. Finally, the structure of a protein can often be segregated into domains that have distinct structures and functions.

2.1 Primary Structure :

The primary structure of a protein refers to the linear sequence of amino acids in the polypeptide chain. The primary structure is held together by covalent bonds such as peptide bonds, which are made during the process of protein biosynthesis or translation. The two ends of the polypeptide chain are referred to as the carboxyl terminus (C-terminus) and the amino terminus (N-terminus) based on the nature of the free group on each extremity. Counting of residues always starts at the N-terminal end (NH2-group), which is the end where the amino group is not involved in a peptide bond. The primary structure of a protein is determined by the gene corresponding to the protein. A specific sequence of nucleotides in DNA is transcribed into mRNA, which is read by the ribosome in a process called translation. The sequence of a protein is unique to that protein, and defines the structure and function of the protein. We know that there are over 10,000 proteins in the human body which are composed of different arrangements of 20 types of amino acid residues.



Copyright © 2008 Pearson Education, Inc., publishing as Pearson Benjamin Cummings.

Fig 2.1: Primary Structure of Protein

2.2 Secondary Structure:

Secondary structure refers to highly regular local sub-structures. There are two main types of secondary structure, the alpha helix and the beta strand or beta sheets. These secondary structures are defined by patterns of hydrogen bonds between the main-chain peptide groups. Both the alpha helix and the beta sheet represent a way of saturating all the hydrogen bond donors and acceptors in the peptide backbone. Some parts of the protein are ordered but do not form any regular structures.

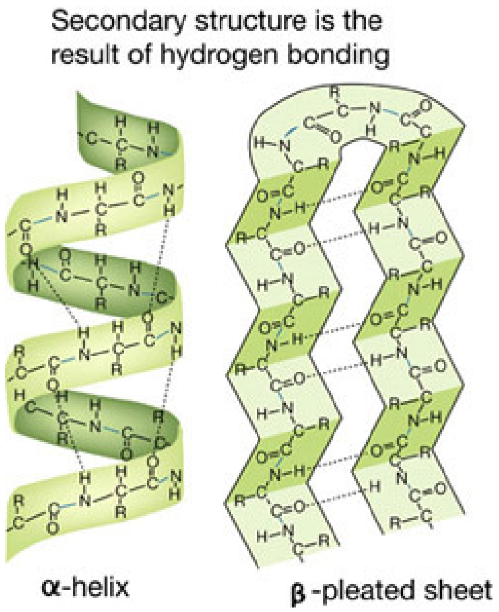


Fig 2.2: Secondary Structure of Protein

2.3 Tertiary Structure:

Tertiary structure refers to the three-dimensional structure of a single, double, or triple bonded protein molecule. The alpha-helixes and beta pleated-sheets are folded into a compact globular structure. The folding is driven by the non-specific hydrophobic interactions, the burial of hydrophobic residues from water, but the structure is stable only when the parts of a protein domain are locked into place by specific tertiary interactions, such as salt bridges, hydrogen bonds, and the tight packing of side chains and disulfide bonds. The disulfide bonds are extremely rare in cytosolic proteins, since the cytosol (intracellular fluid) is generally a reducing environment.

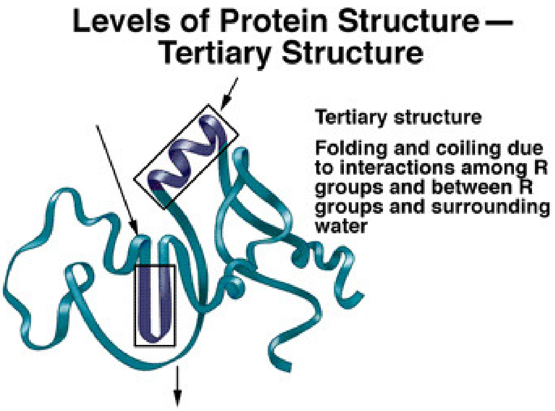

Levels of Protein Structure—
Tertiary Structure
Tertiary structure
Folding and coiling due to interactions among R groups and between R groups and surrounding water

Fig 2.3: Tertiary Structure of Protein

2.4 Quaternary Structure:

Quaternary structure is the three-dimensional structure of a multi-subunit protein and how the subunits fit together. In this context, the quaternary structure is stabilized by the same non-covalent interactions and disulfide bonds as the tertiary structure. Complexes of two or more polypeptides (i.e. multiple subunits) are called multimers. Specifically it would be called a dimer if it contains two subunits, a tri-mer if it contains three subunits, a tetramer if it contains four subunits, and a pentamer if it contains five subunits. The subunits are frequently related to one another by symmetry operations, such as a 2-fold axis in a dimer.

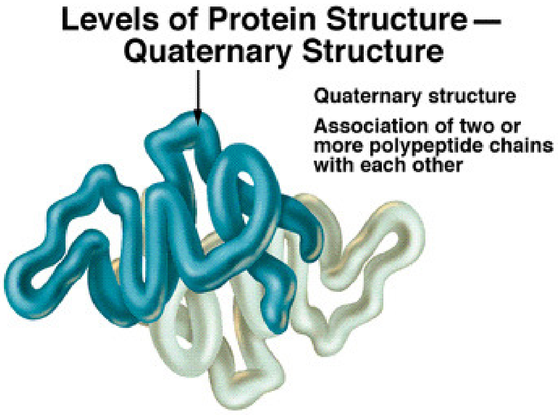

Levels of Protein Structure—
Quaternary Structure
Quaternary structure
Association of two or more polypeptide chains with each other

Fig 2.4: Quaternary Structure of Protein

3. Levinthal Paradox

The original suggestion of a folding pathway which was asserted by C. Levinthal was originally suggested because proteins were found to refold from a disordered state in seconds. Therefore, for a protein to reach its' biologically active, or native state, whether that state exists at a total energy minima or in a metastable local energy minima, it most likely follows a pathway.1 This is because if the protein were to sample random configurations in order to reach its energy minimum, even for a short protein chain of only **~100 amino acids**, it would have to test **2100 ~ 1030** conformations. Therefore for only a picosecond test time of each conformation, it would take **1018 seconds, or over 30 billion years**, which is on the order of the age of our universe, as opposed to the seconds which are taken to fold proteins. Therefore, it can be assumed that due to the speed at which proteins fold, a pathway must exist in order for the protein to fold quickly without error.

4 Kinetic and Thermodynamic Hypothesis

The kinetic and thermodynamic hypotheses were suggested in the original article by C. Levinthal, and postulate on the relative energy of the native state to other possible conformations for the protein.

4.1 Thermodynamic Hypothesis:-

The thermodynamic hypothesis is that the native state exists as the state in normal physiological conditions which minimizes its Gibbs free energy globally. This means that the native state is a unique minimum determined by the protein's amino acid sequence as well as the environment it is in, regardless of the energy bestowed to the protein. Anfinsen et al. provided support for this theory in early work using a mixture of mostly improperly folded protein due to inappropriate environmental conditions (being in 8M urea).The proteins in this environment were only operating at ~1% the activity of the native protein. However, when the protein was removed from the urea and placed in a proper solvent, the same group of proteins eventually formed a homogeneous solution of protein in the native state. This process was determined solely by the free energy minimized by the proteins folding into a more stable energy state, and this state is determined by the amino acid sequence only.

4.2 Kinetic Hypothesis

The kinetic hypothesis is an opposite view, and predicts that the native state is not at a total energy minimum, but is actually at a local minimum which is determined by the folding pathway which achieves a low free energy the quickest. The thermodynamic hypothesis has lots of support, as when it comes to computational work, assumptions are always made that the thermodynamic hypothesis is correct to simplify calculations, yielding results which are usually quite descriptive, and give very strong agreement with experimental findings. However, even though sequence structure of proteins gives accurate results for finding the structure of proteins theoretically at room temperature, there seems to be validity in the kinetic hypothesis.10 As will be discussed later, amyloid fibrils have been shown to be much more stable than the native state of the protein, and thus it is possible that the amyloid fibrils are actually in a more global minimum of free energy, but require extra energy (higher temperatures) to form, supporting the kinetic hypothesis

5. Some Proposed Folding Models

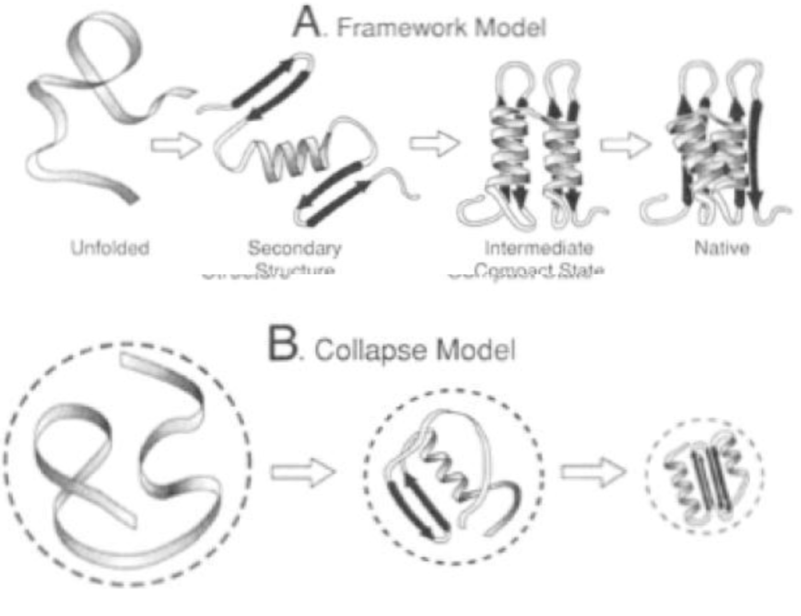
5.1 Hydrophobic Collapse:-

An original consequence of the thermodynamic model was that since the sequence of amino acids in the proteins determines the native state and secondary structure of the protein, the protein folds due to “local” interactions. This means that it forms the secondary structure first, and then arranges itself spatially into its tertiary structure. However, even though proteins are fairly complex molecules, they are still fundamentally polymers. Therefore, it is natural for them in regular physiological conditions to undergo hydrophobic collapse in which they form a globule as opposed to staying as a long stretched chain for an extensive period. This was studied computationally by Sali et al., where they used a Monte Carlo algorithm in order to computationally study the process of folding for a short bead self-avoiding chain.

They used a system where they knew that the native state for the chain would be at a global minimum, in conjunction with the thermodynamic hypothesis. Initially, the particle undergoes a massive hydrophobic collapse which is on a much shorter timescale than the rest of the folding process. Then on a much larger timescale, the particle seems to search for the intermediary transition folding state which will lead it to its native state. This result agrees with the idea of a specific folding pathway for the protein, as the rate limiting step for the folding process to the native state is the search for the intermediary state.

Dill et al. proposed a model which contained the above described properties in opposition to previous models governed by local interactions, forming secondary structures first. He called this the “collapse” model in opposition to Ptitsyn’s “framework” model, and postulated that folding relied on nonlocal instead of local interactions. This collapse model made sense in conjunction with previous physics of polymers, as hydrophobic collapse in proper physiological media is consistent with other polymer conformation and theory. The model he set forth was a “simple exact” model, meaning it approximated each amino acid as a lattice point, or bead, as in the previous study discussed by Sali et al.

They discussed years of experimental evidence which led to the theory that collapse, or non-local interactions, was the dominant force in protein folding. This includes evidence that β-sheet structures had very few local interactions involved in its structure, the free energy difference for helix formation is small, making their formation unstable at physiological temperatures, and the tendency for formation of secondary structures is more dependent on the solvent than on the amino acid sequence. The last result indicated that the local interactions which supposedly caused secondary structure formation initially is not very strong compared to the force imposed by the natural environment. This once again agrees with the well-known drive for polymers to undergo hydrophobic collapse in proper solvents in order to achieve maximum entropy and minimize free energy.



5.1 Multiple Folding Pathways and Intermediary States:-

Studies have shown that proteins may fold along multiple folding pathways in order to change conformation from the random state to native state and vice versa. Therefore instead of a specific pathway leading to the native state, there are many pathways which lead to the folding of the protein in a quick fashion. Originally proposed by Dill et al. was that the energy landscape for the folding of the protein was much like a funnel, hence the name “funnel landscapes”. This idea of funnel landscapes came from the thermodynamic hypothesis which stated the native state was at an energy minimum. Therefore for a protein to fold, it needed to follow a directional path across its energy landscape. Since it had been shown that regardless of the path, the protein in the proper conditions would eventually find its way from the unfolded state to the native state, the energy landscape would take on a funnel shape, where the folding would be inevitably directional toward its energy minimum.

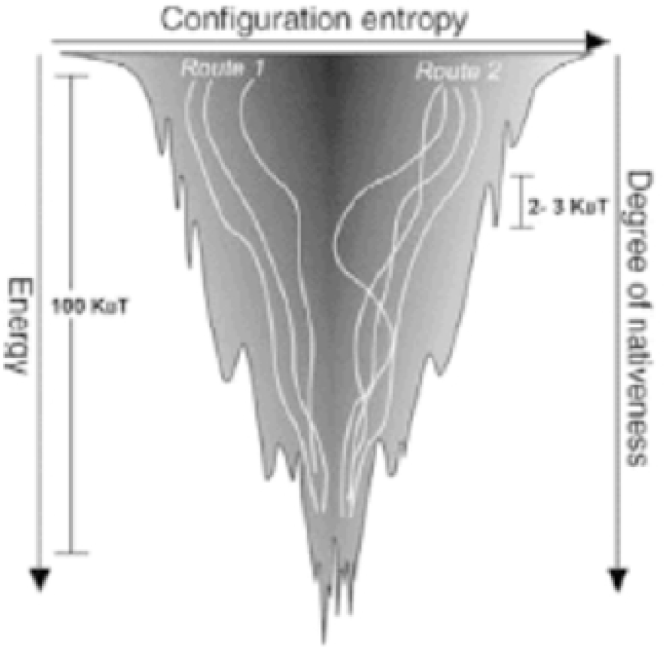
Rhoades et al. studied single adenylate kinase protein molecules which were placed in conditions favouring spontaneous transitions between folded and unfolded states as found in previous ensemble measurements of adenylate kinase.15 Single molecule studies are a fairly recent possibility in protein folding.15,19 In this particular study, they used a technique called Fluorescence Resonance Energy Transfer (FRET) which features a donor dye attached to one end of the protein chain and an acceptor dye to the other. The protein is illuminated with laser light which excites the donor dye. When the donor dye is close enough to the acceptor dye, there is energy transfer without radiation to the acceptor dye, which is then excited and emits radiation of a separate wavelength. This is dependent on the acceptor dye having an excitation energy in the range of the emission energy for the donor dye. By measuring the intensity of fluorescence emitted by the donor and acceptor dye, we define an efficiency of energy transfer as follows:-

$$E_{ET} = \frac{I_a}{I_a + I_d}$$

where I_a is the acceptor emission intensity detected, and *Id* is the donor emission intensity detected, we can determine how folded the molecule is, as the energy transfer from donor to acceptor is extremely sensitive to distance of separation:

$$E \sim \frac{R_0^6}{R_0^6 + r^6}$$

where *R0* is the Forster's radius, usually on the order of 1-5 *nm*, and *r* is the distance of separation for the donor and acceptor dyes. It is worth noting here that for high efficiency, the distance between dyes must be small, and thus the protein is in a folded form.



5.3 Two State Folding

Since protein folding is complex, and there is evidence of multiple folding pathways, it is very difficult to analyze and interpret protein folding data. Therefore in order to obtain quantitative understanding of protein folding, groups have studied the two state folding model in conjunction with short stranded proteins.

The two state folding model relies on the fact that the protein may exist in only two states: folded (F) and unfolded (U). The model is set up much like a chemical reaction, with a before and after state, with folding and unfolding rates. Thus we can describe the kinetics of folding as well as the equilibrium transitions in a straightforward method. Essentially, the model treats the two states as a reaction with an activation barrier. Therefore, the equilibrium condition can be easily seen to be from elementary chemistry:

$$\frac{[f]}{[u]} = e^{-(E_f - E_u)/kT}$$

where E_f and E_u are the energy barriers for folding and unfolding respectively, k is Boltzmann's constant, and T is the temperature of the system.

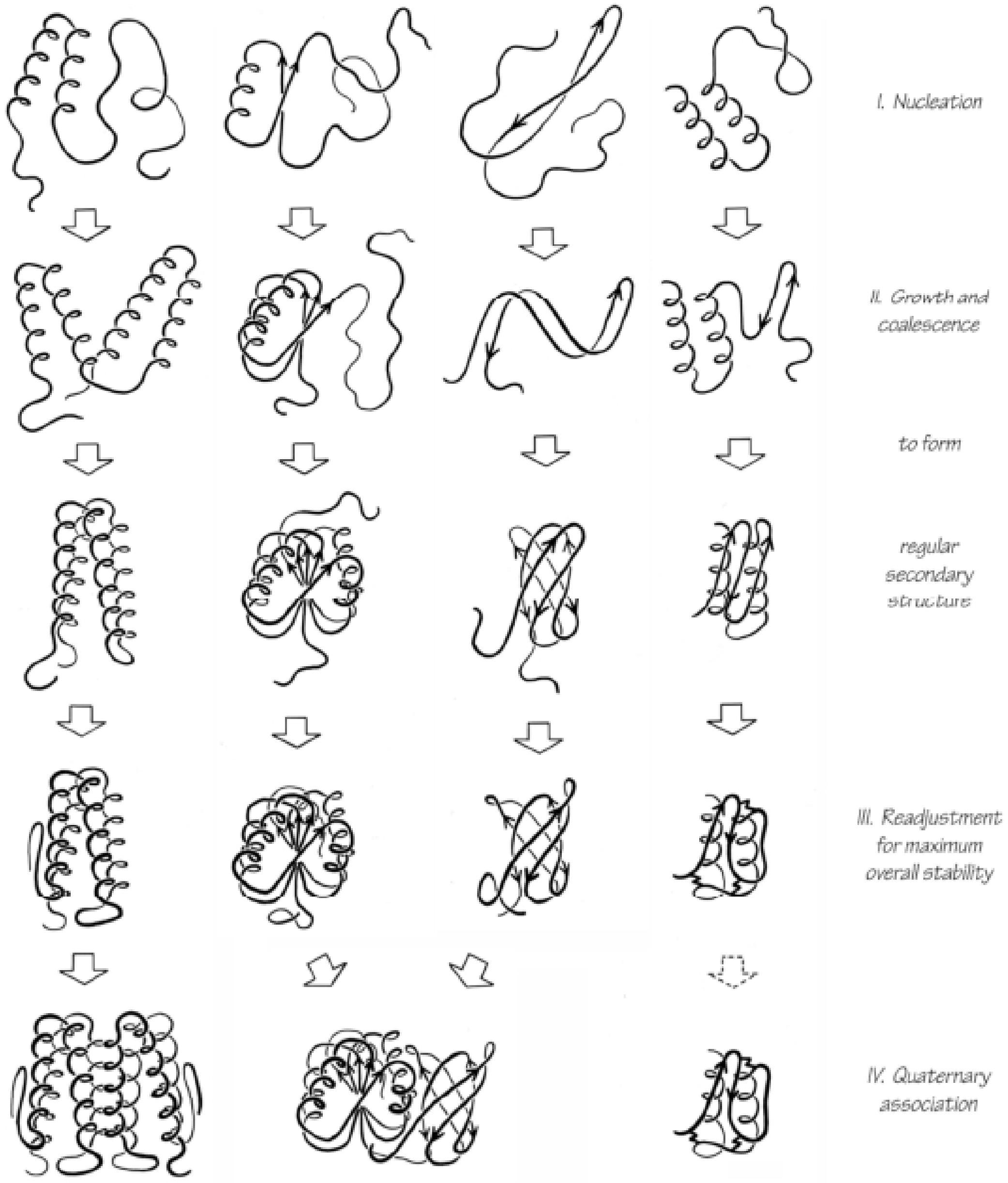
The kinetics may also be studied, as the rates of folding and unfolding are:

$$k_f = Ae^{-E_f/kT}$$

$$k_u = Ae^{-E_u/kT}$$

where k_f and k_u are the folding and unfolding rates respectively, and A is just a constant. Using these kinetic rates, groups may obtain a quantitative understanding of the energy barrier for folding of the small proteins. Yang et al. carried out a study which used an ensemble of a small protein, meaning a large amount of them were studied at the same time. The experiment consisted of perturbing a large group of protein molecules in their native state with a sudden increase in temperature. This temperature increase forced the proteins to unfold, and when the solution was placed back at physiological temperature, the relaxation rate of the proteins back into the folded state was measured. By taking the ratio of the unfolding rate to the folding rate, they were able to determine the activation barrier over which the folding occurs for the proteins. Figure 9 shows an example of the exponential relaxation of the particle from the unfolded to the folded state, and gives the average amount of time a particle spends unfolded when placed in proper folding conditions. This example gives a rate of ~0.05 μs^{-1} and indicates an activation barrier on the order of $2.1kT$.23

The two state model is a valid method for evaluating quantitatively the folding of small, fast folding proteins. Although, in reality you are almost guaranteed to see intermediate folding steps if you decrease your detection time enough. The model does however give us a way to quantify the folding process, and ignore the complexity of the other models for folding. This simplicity is desirable in a field of research where complexity is far too apparent.


I. Nucleation
II. Growth and coalescence
to form
regular secondary structure
III. Readjustment for maximum overall stability
IV. Quaternary association

6. Power Laws in Protein Folding

6.1 Two State Folding

The correlation function for a field φ(x) is defined as:-

$$G(x,y) = \langle \phi(x)\, \phi(y) \rangle - \langle \phi(x) \rangle \langle \phi(y) \rangle$$

Generally, $G(x, y)$ decreases with increasing distance $|x - y|$ exponentially, for large distances:

$$G(x,y) \sim e^{-|x-y|/\xi}$$

The characteristic length ξ is called the *correlation length*. We shall show that it diverges at the critical point.

6.2 Universality Classes:-

The correlation length ξ measures the distance within which values of the field are correlated. We cannot resolve spatial structures smaller than ξ, because the field organizes itself into uniform blocks of approximately that size. As we approach the critical point, increases, and we lose resolution. At the critical point, when ξ diverges, we cannot see any details at all. Only global properties, such as the dimension of space, or the number of degrees of freedom, distinguish one system from another. That is why systems at the critical point fall into universality classes, which share the same set of critical exponents.

6.3 Compactness Index:-

For a distribution of N points set up by a given rule, the characteristic length R of the system depends on N, as $N \to \infty$, through a power law:

$$|R = aN^{\nu}$$

where a is a length scale. The exponent ν is an indication of the compactness of the system. For example:

(i) For ordinary matter ν = 1⁄3, since it “saturates,” i.e. the density N/R^3 is constant.

(ii) For Brownian motion ν = 1⁄2, since the average end-to-end distance for N steps is proportional to $\sqrt{N}$.

If we consider the correlation function $g(r)$ of a protein, which gives the probability of finding a residue at r, when it is known that one is at the origin. We envision the limit in which the number of residues and the average radius tend to infinity. Let η be the scale of atomic distances in the molecule, and ξ the correlation length.

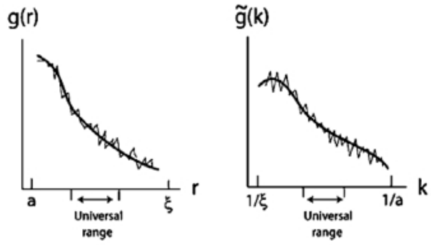
On small length scales, $g(r)$ has fluctuations over distances of order η. We shall smooth out such fluctuations through spatial averaging. In experiments, the averaging is done through the finite resolution of the measuring instruments.

On large scale length g(r) tends to 0 for r>> ξ

The *universal range* lies in between, where $g(r)$ obeys a power law:

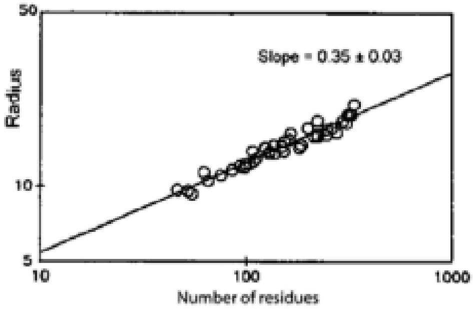
$$\eta \ll r \ll \xi$$
$$\eta^{-1} \gg k \gg \xi^{-1}$$

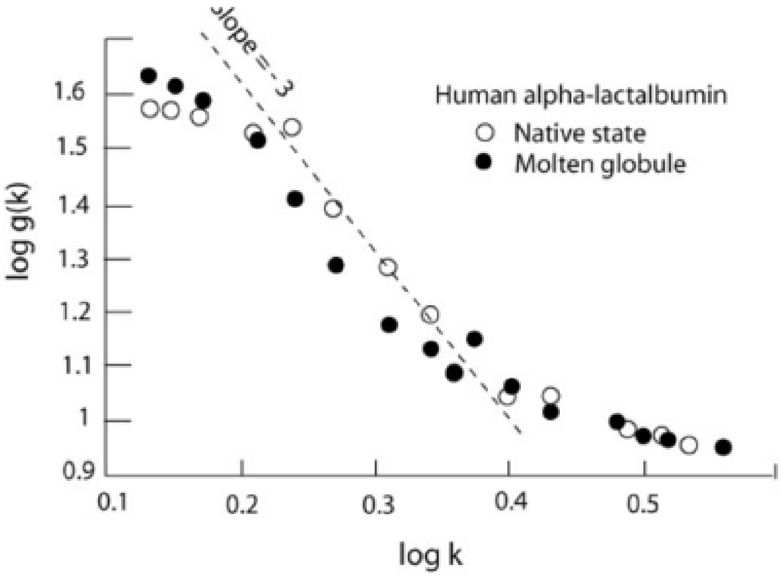
Since r <<ξ, we are in the critical region where detailed structures of the system become irrelevant.



6.4 Collapse and Annealing:-

The native protein state is as compact as ordinary matter, with a characteristic density. This corresponds to a compactness index ν = 1⁄3, as shown in the plot of empirical data1 in Fig. 14.2. Thus, in the universal range, the native state exhibits the power law.

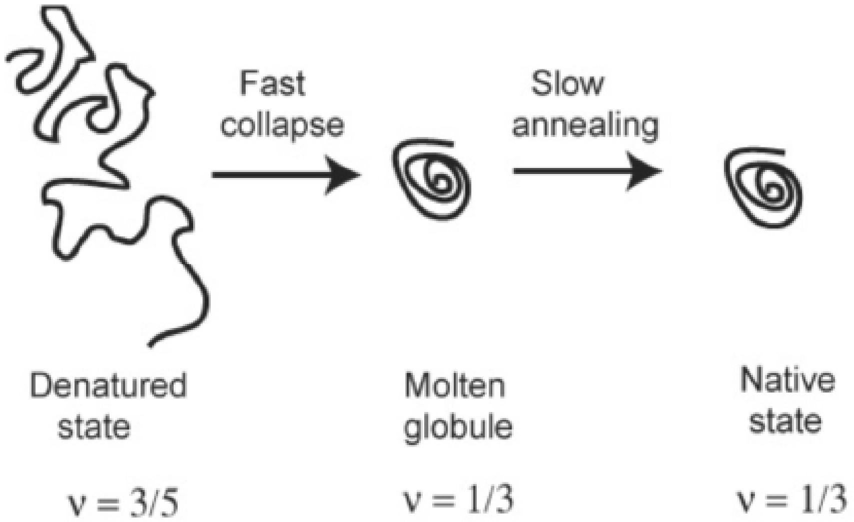




In the denatured state, on the other hand, the protein is a random coil. In the universal range, where local structures are irrelevant, the structure is similar to a random polymer chain, for which the compactness index is well approximated by ν = 3/5, and thus:-

$$g(k) \sim k\ -5/3 \text{ (Denatured state)}$$

This shows that the molten globule is as compact as the native state.



Fast collapse and slow annealing in protein folding. The compactness index ν changes from 3/5 to 1/3, the latter being the same as that for ordinary matter. In summary, the folding process consists of a fast collapse, followed by slow annealing

7. Self-Avoiding Walk

7.1 Self Avoiding Walk and Compactness Index

In mathematics, a self-avoiding walk (SAW) is a sequence of moves on a lattice that does not visit the same point more than once. This is a special case of the graph theoretical notion of a path. SAWs were first introduced in order to model the real-life behavior of chain-like entities such as solvents and polymers, whose physical volume prohibits multiple occupation of the same spatial point.

It is easy to generate a sequence of SAW steps on a computer. A straightforward algorithm is to first generate a random step, and accept it if an old site is not revisited, and generate another step otherwise. One repeats the process until N acceptable steps are obtained, and this represents one state of a polymer chain. By generating a sufficient number of chains independently, one has an ensemble of polymer chains, in which statistical averages can be calculated. It is harder to simulate the dynamics — the Brownian motion of the chain. The most successful algorithm appears to be the “pivoting method,” in which one rotates part of a chain about an arbitrarily chosen monomer, and accept the outcome only if it is self-avoiding.

The compactness index can be calculated approximately, using an intuitive argument due to Flory. Consider a chain of N monomers, with end-to-end distance R in D-dimensional space, in the limit N →∞, R→∞. The monomer density is given by

$$n \sim \frac{N}{R^D}$$

The energy of the chain should contain a repulsive term due to self-avoidance, with energy density ε proportional to n 2 , the number of pairs per unit volume:

$$\varepsilon \sim kTv_0n^2$$

where v_0 is the excluded volume, and kT is the thermal energy. The total repulsive energy is

$$E_{\text{repulsive}} \sim \varepsilon R^D \sim \frac{kTv_0N^2}{R^D}$$

The repulsive energy tends to expand the volume at small R. As R increases, however, the collision between two monomers becomes increasingly rare, and the sequence of steps should approach a simple random walk. This effect is taken into account by postulating a linear restoring force, corresponding to an "elastic energy" proportional to R^2 :

$$E_{\text{elastic}} \sim k_BT\left(\frac{R}{R_0}\right)^2 \sim \frac{kTR^2}{Na^2}$$

where R_0 = a √ N is the characteristic length in simple random walk.

The total energy is of the form:

$$E = E_{\text{repulsive}} + E_{\text{elastic}} \sim kT\left(\frac{v_0N^2}{R^D} + \frac{R^2}{Na^2}\right)$$

Minimizing this with respect to R, we obtain the equilibrium radius

$$R_{\text{eq}} \sim N^{3/(D+2)}$$

The compactness index is therefore:

$$\nu = \frac{3}{D+2}$$

7.2 Self Avoiding Walk Simulation

We wrote a Java program to simulate a self-avoiding random walk on a 2-D surface. Self-avoiding random walks arise in modeling physical processes like the folding of polymer molecules. Such walks are difficult to model using classical mathematics. So they are best studied by direct numerical simulation.

The program starts at N = 10 and gathers data for 50 runs. In each run, the program simulates a self-avoiding random-walk for the given N points. The data is then processed to calculate the root-mean-square end to end distance, R. The sample simulation was run till N = 100. The data for each run is non deterministic in the sense that there is no relation between the path chosen for an arbitrary run and any other run. Every run is random but may or may not be unique. The aim of the simulation was to find the compactness index and verify it with Flory's prediction (previous page).

From the previous section,

$$R_{\text{eq}} \sim N^{3/(D+2)}$$

Where, compactness index,

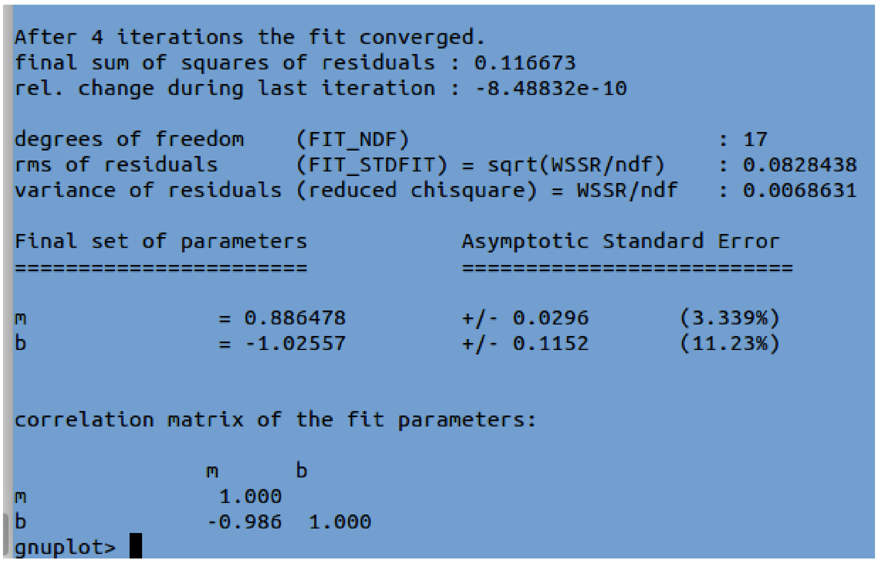
$$\nu = \frac{3}{D+2}$$

For a 2-D simulation, D = 2, hence expected **v = ¾ = 0.75.** Now,

$$\log (R_{eq}) = 0.75 \log(N) + \log (c)$$

Hence, a graph between log (R_{eq}) and log (N) should give slope, m = 0.75, which is the aim of the simulation.

The following are the results of the simulation:



```
After 4 iterations the fit converged.
final sum of squares of residuals : 0.116673
rel. change during last iteration : -8.48832e-10

degrees of freedom    (FIT_NDF)                        : 17
rms of residuals      (FIT_STDFIT) = sqrt(WSSR/ndf)    : 0.0828438
variance of residuals (reduced chisquare) = WSSR/ndf   : 0.0068631

Final set of parameters            Asymptotic Standard Error
=======================            ==========================

m               = 0.886478         +/- 0.0296       (3.339%)
b               = -1.02557         +/- 0.1152       (11.23%)


correlation matrix of the fit parameters:

               m      b
m              1.000
b             -0.986  1.000
gnuplot> █
```

Fig 4.1: A best fit line of the simulation data gives m = 0.886 +/- 11.23%.

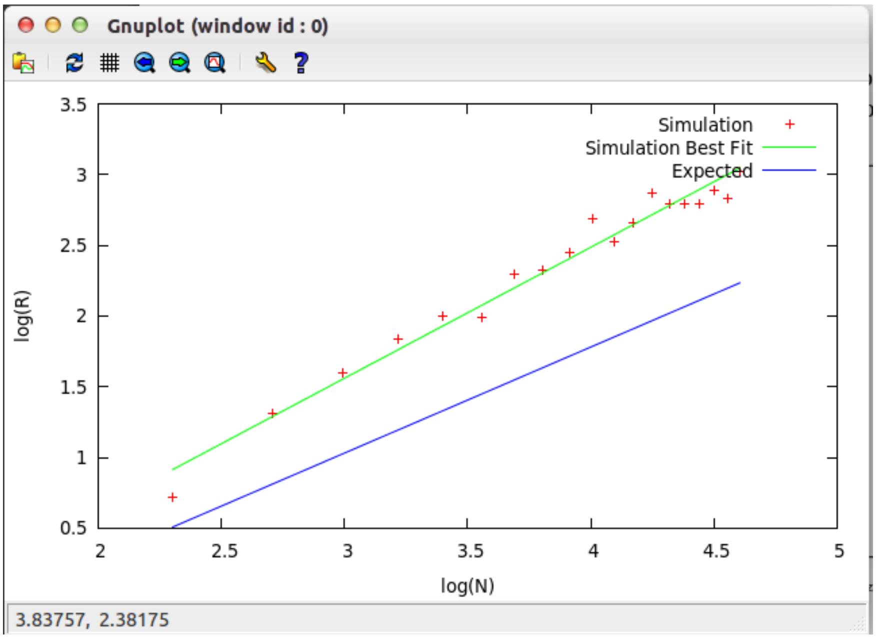


Fig 4.2: Each Simulation point was calculated as t

The Simulation gives Compactness Index as **0.886 +/- 11.23%** as compared to the calculated value of 0.75.

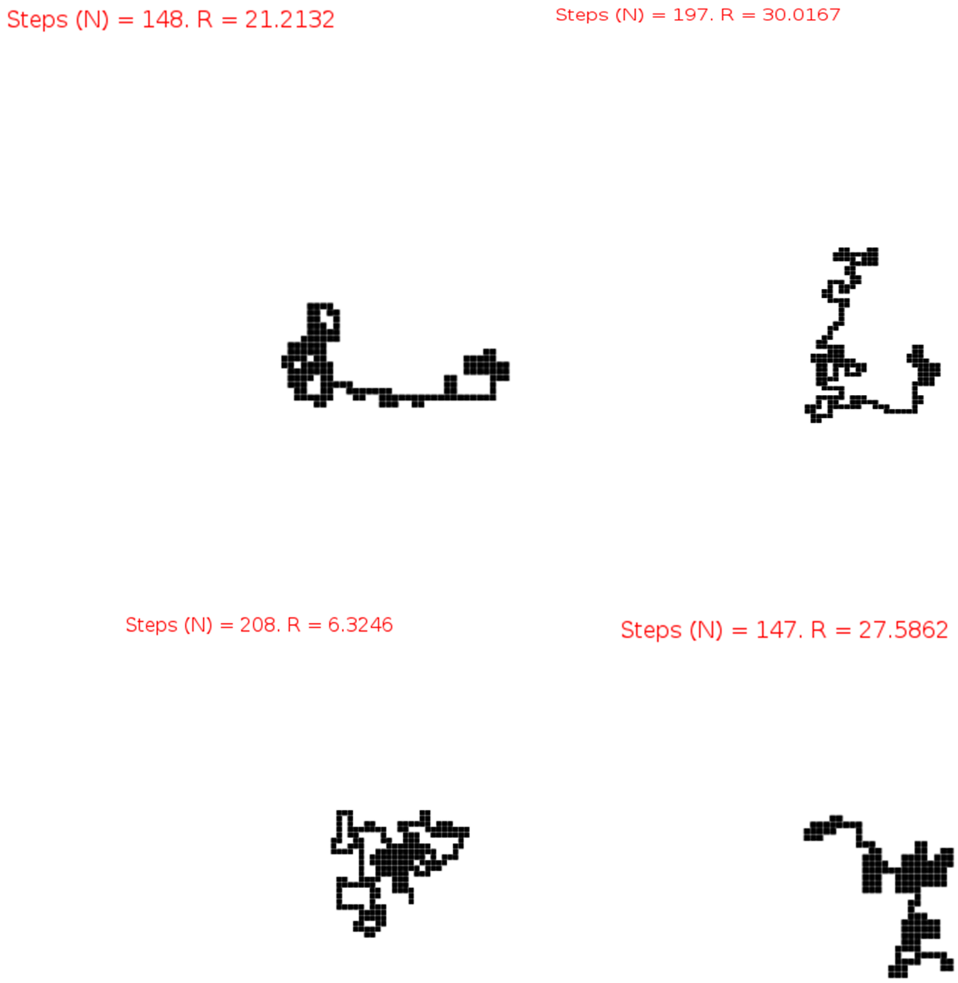

Steps (N) = 148. R = 21.2132
Steps (N) = 197. R = 30.0167
Steps (N) = 208. R = 6.3246
Steps (N) = 147. R = 27.5862

Fig 4.3: Different runs of a self-avoiding random walk with the same parameters.

8. QCD and Protein Dynamics

From the biochemical point of view, proteins are poly-peptide chains, polymers formed by 20 different types of monomers called amino-acids. Typical globular proteins consist of about 100 amino-acids, but some chains can contain as many as ~ 30,000 of such monomers.

From the physical point of view, proteins are very different from essentially all other known polymers: indeed, while standard hetero-polymers are prototypes of disordered mesoscale systems, proteins are characterized by their unique ability to spontaneously and reversibly fold into a well-defined stable configuration (denominated the native state), in which they perform their biological functions. Such a remarkable feature is the result of the evolutionary pressure, which has selected only a very specific subset in the space of all possible amino-acid sequences.

The native state of a protein is characterized by an extremely low entropy, since it is defined by the small thermal oscillations around a single conformation. In this sense proteins may be regarded as non-symmetric mesoscale crystals. If the temperature of the water surrounding the the solvent, then proteins can reach a much more entropic phase called the molten globule. In this state, the chain is still collapsed, but is free to explore the different compact configurations, in analogy with a liquid drop. Finally, at even higher temperatures or denaturant concentrations, proteins swell and start to behave as flexible random coils.

Interestingly, the phase diagram of proteins in solution displays several analogies with that of strongly interacting matter. Indeed, the low-temperature and low-density phase of QCD is characterized by a very low only colorless hadrons entropy, since only colorless hadrons are present in the spectrum. On the other hand, above the de-confinement phase transition, the entropy is much larger, since also quantum states in color multiplets can contribute to the partition function.

8.1 Path Integral Formulation of Molecular Dynamics

The dynamics of bio-molecular systems in solution is often described by means of the Langevin equation, which reads

$$m\;\ddot{x} = -\gamma\dot{x} - \nabla U(x) + \xi(t),$$

In this equation, x = (x 1 ,x 2 ,...,x N) is a macro-vector specifying the coordinates of all the atomic nuclei in the molecule, U(x) is the potential energy and ξ(t) is a stochastic force, which simulates the effects of the random Brownian collisions with the molecules in the solvent. The contemporary models for the potential energy U(x)

For sake of simplicity and without loss of generality, in the following we present the formalism in the case of one particle in one spacial dimension. The generalization to the multi-dimensional case is straightforward. The acceleration term on the left hand side of Eq. introduces inertial effects which are damped on a time-scale t ~ m/γ ≡ τ D. For most biomolecular systems, τ D is of order of the fraction of a ps, hence much smaller than the relevant microscopic time-scale associated to the dynamics of the torsional angles, which takes place at the ns time-scale.

For t >> τ D the Langevin Eq. reduces to

$$\frac{\partial x}{\partial t} = -\frac{D}{k_B T} \nabla U(x) + \eta(t)$$

where D is the so-called diffusion coefficient and η(t) is a rescaled white noise that satisfy the so-called fluctuation-dissipation relationship:

$$\langle \eta(t)\eta(t') \rangle = 2D\delta(t-t').$$

It can be shown that the probability distribution sampled by the stochastic differential Eq. satisfies then the Fokker-Planck (FP) equation:

$$\frac{\partial}{\partial t} P(x,t) = D\nabla \left(\frac{1}{k_B T} \nabla U(x)\ P(x,t) \right) + D\,\nabla^2\ P(x,t).$$

A universal property of all the solutions of the Eq. (4) is that, in the long time limit, they converge to the Boltzmann weight, regardless of their initial conditions.

By performing the substitution,

$$P(x,t) \equiv e^{-\frac{1}{2k_B T}U(x)}\ \psi(x,t).$$

the FP Eq. can be formally re-written as a Schrödinger Equation in imaginary time:

$$-\frac{\partial}{\partial t}\psi(x,t) = \hat{H}_{eff}\ \psi(x,t),$$

where the effective ”Quantum Hamiltonian” operator reads

$$\hat{H}_{eff} \;=\; -D\nabla^2 + V_{eff}(x).$$

and V$_{eff}$ (x) is an effective potential and is defined as:

$$V_{eff}(x) = \frac{D}{4(k_B T)^2}\left[(\nabla U(x))^2 - 2k_B T\,\nabla^2 U(x)\right]$$

Hence, we have shown that the problem of studying the real time stochastic dynamics of a protein in solution can be mapped into the problem of determining the imaginary time evolution of an effective quantum many-body system.

Based on the analogy with quantum mechanics, it is immediate to obtain a path-integral representation of the solution of (4), subject to the boundary conditions x(0) = x$_i$ and x(t) = x$_f$

$$\begin{aligned} P\left(x_f, t\,|x_i\right) &= e^{-\frac{1}{2k_BT}(U(x_f)-U(x_i))}\langle x_i|e^{-\hat{H}_{eff}t}|x_f\rangle \\ &= e^{-\frac{1}{2k_BT}(U(x_f)-U(x_i))}\int_{x_i}^{x_f}\mathcal{D}x(\tau)\; e^{-\int_0^t d\tau\,\left(\frac{\dot{x}^2(\tau)}{4D}+V_{eff}[x(\tau)]\right)} \end{aligned}$$

This equation expresses the fact that the Green’s function of the Fokker-Planck equation is formally equivalent to a quantum-mechanical propagator in imaginary time.

9. Two State Kinetic Model

The folding of some proteins appears to be a two-state kinetic process. A two-state kinetic model is justified if protein molecules rapidly equilibrate between different unfolded conformations prior to complete folding.

The rate equation is as follows:-

$$\frac{dP_N}{dt} = k_f P_U - k_u P_N, \qquad P_N + P_U = 1,$$

Where, P_N is the fraction of protein in its native state *N*, P_U is the fraction of protein in the unfolded state *U*, The folding rate is k_f and the unfolding rate is k_u.

State denotes a region of configuration space, usually the neighborhood of a potential minimum. The native state is associated with the deepest minimum. The ‘‘unfolded state’’ is the rest of configuration space. It is made up of a large number of distinct regions, each one associated with a local minimum or conformation of the polypeptide chain.

If the regions of configuration space are properly chosen, the protein remains in one state long enough to reach local equilibrium, and then jumps to another state. Thus folding appears to be intrinsically a many-state kinetic process, described by the more general rate equation:-

$$\frac{dP_a}{dt} = \sum_b k(b \rightarrow a)P_b - \sum_b k(a \rightarrow b)P_a,$$

where *Pa* is the fraction of protein in a particular region or conformational state *a*, the native state is *a* = *N*, and the ‘‘unfolded state’’ is all *a* not equal to *N*.

The first sum is the total gain in state *a* due to transitions from other states *b*, and the second sum is the total loss from state *a* due to transitions to other states *b*. The long time limit P_a (*t* tending to infinity) of the solution of these equations is the equilibrium fraction P_a(eq).

Under folding conditions, unfolded protein molecules rapidly equilibrate between different conformations prior to complete refolding. In this view, the ‘‘unfolded state’’ is actually an equilibrium distribution of many unfolded or partially folded conformational states.

We show here that this rapid equilibration is a consequence of reasonable assumptions about rate constants and the thermodynamics of folding.

9.1 Rate Constants

The rate constant *k*(*a to b*) has a special structure which leads to rapid equilibration of the unfolded states. It is determined by a quantity $\boldsymbol{B_{a,b}}$ that depends only on the boundary dividing the initial state *a* and the final state *b*, and by **the partition function $\boldsymbol{Q_a}$** of the initial state,

$$k(a \rightarrow b) = \frac{B_{a,b}}{Q_a}.$$

The boundary factor is the same whether the protein moves from *a* to *b* or from *b* to *a*,

$$B_{a,b} = B_{b,a},$$

The reaction rate constants must satisfy the principle of detailed balance; at equilibrium, forward and backward rates must be equal,

$$P_a(\mathrm{eq})k(a \rightarrow b) = P_b(\mathrm{eq})k(b \rightarrow a).$$

This condition is automatically satisfied, because P_a(eq) is proportional to the partition function Q_a that appears in the denominator of k(a to b).

9.2 Kinetic Theory and Folding Thermodynamics:-

Each individual state a has its own partition function Q_a. The native state has the partition function Q_N. The purpose of the present section is to show that under folding conditions, Q_a for any single unfolded state is much smaller than Q_N.

The total partition function Q_{total} is the sum over all states,

$$Q_{\text{total}} = \sum_a Q_a.$$

The equilibrium fraction of protein in the single state a is,

$$P_a(\text{eq}) = \frac{Q_a}{Q_{\text{total}}}.$$

In particular, the fraction of protein in the native state $a = N$ is,

$$P_N(\text{eq}) = \frac{Q_N}{Q_{\text{total}}}.$$

The folding transition is produced by a change in folding conditions, for example in temperature or in the concentration of a denaturant. If the folding conditions favor a completely unfolded protein, all of the partition functions Q_a are roughly the same size and all of the rate constants are comparable. Suppose that experimental conditions are changed to favor folding, so that the equilibrium fraction of native protein varies from small, say *PN* (eq) approx. 0.01, to large, *PN* (eq) approx. 1. . Then in this range of folding conditions, the partition function of the native state is a substantial fraction of the total partition function, $\boldsymbol{Q_N/Q_{\text{total}}>0.01}$. The remaining partition function of all the unfolded states,

$$Q_U = \sum_{a \neq N} Q_a,$$

This quantity is always smaller than Q total, and the ratio Q_U/Q_N is always smaller than 100. The ratio of the partition function Q_a of a single unfolded state to the partition function *QN* of the native state is limited by

$$\frac{Q_a}{Q_N} = \frac{Q_a}{Q_U}\frac{Q_U}{Q_N} < 100\,\frac{Q_a}{Q_U}.$$

If there are a very large number of unfolded states, each one can make only a small contribution to Q_U, and the ratio Q_a/Q_U is expected to be very much smaller than 1/100. Under folding conditions, any individual Q_a is much smaller than Q_N.

If the protein is in a gateway state, it can make transitions to the single native state or to many other unfolded states, all with comparable rates. Then transitions between unfolded states are statistically more likely than transitions into the native state. This observation, along with the earlier estimate of the relative order of magnitude of rate constants, leads to the conclusion that folding kinetics involves two distinctly different time scales.

The fast time scale extends from microseconds to milliseconds, and the slow time scale may require seconds or minutes. In the fast time scale, the unfolded protein moves rapidly between unfolded or partially folded conformational states. After a short time these states come to local thermodynamic equilibrium, and all details about the initial state and the sequence of transitions (the ‘‘folding mechanism’’) are forgotten. The fraction of folded protein varies on the slow time scale.

Now it is easy to show explicitly how the many-state rate equation reduces to the two-state equation. Transitions between unfolded states are fast, and the ensemble of unfolded states relaxes rapidly to local thermodynamic equilibrium. However, after this fast relaxation the total fraction of unfolded states will still change, as the fraction in the native state changes. This can be handled by a time-dependent normalization coefficient $c(t)$, so that after the fast relaxation, the time-dependent $Pa(t)$ is proportional to the equilibrium $Pa(\text{eq})$,

$$P_a(t) \rightarrow c(t)P_a(\text{eq}).$$

The coefficient is determined by a normalization condition,

$$\sum_{a\neq N} P_a(t) = 1 - P_N(t).$$

This is the approximation that leads to two-state kinetics. When it is inserted into the rate equation for the native state,

$$\frac{dP_N(t)}{dt} = \sum_{a\neq N} k(a \rightarrow N)P_a(t) - \sum_{a\neq N} k(N \rightarrow a)P_N(t),$$

The result is,

$$\frac{dP_N(t)}{dt} = \sum_{a\neq N} k(a \rightarrow N)\frac{1-P_N(t)}{1-P_N(\text{eq})}P_a(\text{eq})$$
$$-\sum_{a\neq N} k(N \rightarrow a)P_N(t).$$

The first term can be simplified by using the detailed balance condition:

$$k(a \rightarrow N)P_a(\text{eq}) = k(N \rightarrow a)P_N(\text{eq}).$$

$$\frac{dP_N(t)}{dt} = \sum_{a\neq N} k(N \rightarrow a)\frac{1-P_N(t)}{1-P_N(\text{eq})}P_N(\text{eq})$$
$$-\sum_{a\neq N} k(N \rightarrow a)P_N(t).$$

This has exactly the structure of the two-state kinetic model,

$$\frac{dP_N(t)}{dt} = k_{\mathrm{f}}(1 - P_N(t)) - k_{\mathrm{u}}P_N(t).$$

The unfolding rate is the sum of all transition rates from the native state to all gateway states,

$$k_{\mathrm{u}} = \sum_{a \neq N} k(N \rightarrow a)$$

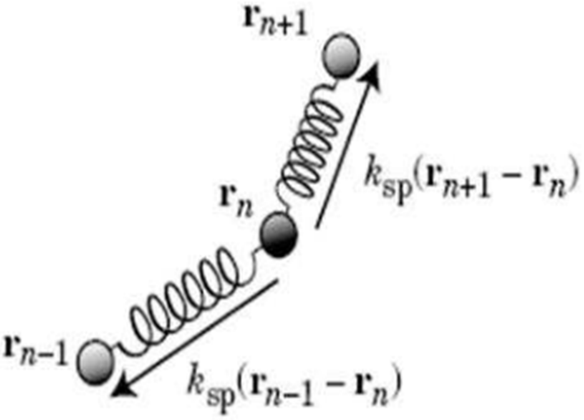
and the folding rate is,

$$k_{\mathrm{f}} = k_{\mathrm{u}} \frac{P_N(\mathrm{eq})}{1 - P_N(\mathrm{eq})}.$$

This qualitative justification of a two-state kinetic model was based on a number of assumptions. (*i*) The statistical thermodynamics of the folding transition is well described by a single folded state and a large ensemble of unfolded states. (*ii*) Any individual unfolded state makes a very small contribution to the total partition function of unfolded states. (*iii*) Transition rate constants are inversely proportional to the partition functions of single conformational states, and unfolded states make transitions to many other unfolded states. (*iv*) The boundary factors in rate constants are insensitive to changes in folding conditions. (*v*) There are no exceptional barriers between particular unfolded states that might lead to trapping for long periods of time. If a particular protein meets these requirements, then one should expect that its folding kinetics is well described by a two-state model.

10. Rouse Model for Polymer Dynamics

The first successful molecular model of polymer dynamics was developed by Rouse. The chain in the Rouse model is represented as N beads connected by springs of root- mean-square size b, as shown in Figure 2.1. The beads in Rouse model only interact with each other through the connecting springs. Each bead is characterized by its own independent friction according to Stokes law with the friction coefficient ζ = 6πηR. Solvent is assumed to be freely draining through the chain as it moves. Detailed mathematical formulation is given below.


$\mathbf{r}_{n+1}$
$\mathbf{r}_n$
$k_{sp}(\mathbf{r}_{n+1} - \mathbf{r}_n)$
$\mathbf{r}_{n-1}$
$k_{sp}(\mathbf{r}_{n-1} - \mathbf{r}_n)$

Interactions between monomers in Rouse model.

10.1 Basic Formulation

Let (R1, R2,..., RN) ≡ {RN } be the positions of the beads connected along a polymer chain. They will satisfy the following Langevin equation (as discussed in section 5.1).

$$\frac{d\mathbf{R}_N}{dt} = -\frac{1}{\zeta}\frac{\partial U}{\partial \mathbf{R}_N} + \mathbf{f}_n$$

The interaction potential are written as:

$$U = \frac{k}{2}\sum_{n=2}^{N}(\mathbf{R}_n - \mathbf{R}_{n-1})^2 \quad with \quad k = \frac{3k_BT}{b^2}$$

For internal beads (n = 2,3,...,N − 1),

$$\zeta\frac{d\mathbf{R}_n}{dt} = -k(2\mathbf{R}_n - \mathbf{R}_{n-1} - \mathbf{R}_{n+1}) + \mathbf{f}_n$$

For the end beads (n = 1 and N)

$$\zeta\frac{d\mathbf{R}_1}{dt} = -k(\mathbf{R}_1 - \mathbf{R}_2) + \mathbf{f}_1, \quad \zeta\frac{d\mathbf{R}_N}{dt} = -k(\mathbf{R}_N - \mathbf{R}_{N-1}) + \mathbf{f}_N$$

We re-define the terminal beads as:

$$\mathbf{R}_0 = \mathbf{R}_1, \quad \mathbf{R}_{N+1} = \mathbf{R}_N$$

In the continuous limit,

$$\zeta\frac{\partial \mathbf{R}_n}{\partial t} = -k\frac{\partial^2 \mathbf{R}_n}{\partial n^2} + \mathbf{f}_n$$

$$\left.\frac{\partial \mathbf{R}_n}{\partial n}\right|_{n=0} = 0, \quad \left.\frac{\partial \mathbf{R}_n}{\partial n}\right|_{n=N} = 0$$

The distribution of random force f n is Gaussian, characterized by,

$$< \mathbf{f}_n(t) >= 0$$

$$< f_{n\alpha}(t) f_{m\beta}(t') >= 2\zeta k_B T \delta_{nm}\delta_{\alpha\beta}\delta(t - t')$$

10.2 Solution in Normal Coordinate

As shown above, the Langevin equation in the continuous limit, has the form of a linear harmonic oscillator, so we introduce following normalized coordinates:

$$\mathbf{X}_p \equiv \frac{1}{N}\int_0^N dn \cos(\frac{p\pi n}{N})\mathbf{R}_n(t) \quad with \quad p = 0, 1, 2\ldots$$

The inverse transform is:

$$\mathbf{R}_n = \mathbf{X}_0 + 2\sum_{p=1}\mathbf{X}_p \cos(\frac{p\pi n}{N})$$

The Langevin equation can also be written as:

$$\zeta_p \frac{\partial \mathbf{X}_p}{\partial t} = -k_p \mathbf{X}_p + \mathbf{f}_p$$

Where,

$$\zeta_0 = N\zeta \quad and \quad \zeta_p = 2N\zeta \quad for \quad p = 1, 2, \ldots$$
$$k_p = \frac{2\pi^2 k p^2}{N} = \frac{6\pi^2 k_B T}{Nb^2}p^2 \quad for \quad p = 0, 1, 2, \ldots$$

and the f_p are the random forces which satisfy:

$$< \mathbf{f}_p(t) >= 0$$
$$< f_{p\alpha}(t) f_{q\beta}(t') >= 2\zeta_p k_B T \delta_{pq}\delta_{\alpha\beta}\delta(t - t')$$

For p > 0,

$$< \mathbf{X}_{p\alpha}(t)\mathbf{X}_{q\beta}(0) >= \delta_{pq}\delta_{\alpha\beta}\frac{k_B T}{k_p}\exp(-t/\tau_p)$$

Where,

$$\tau_1 = \frac{\zeta_1}{k_1} = \frac{\zeta N^2 b^2}{3\pi^2 k_B T} \quad and \quad \tau_p = \frac{\tau_1}{p^2}$$

For p = 0,

$$< (\mathbf{X}_{0\alpha}(t) - \mathbf{X}_{0\alpha}(0))(\mathbf{X}_{0\beta}(t) - \mathbf{X}_{0\beta}(0)) >= \delta_{\alpha\beta}\frac{2k_BT}{\zeta_0}t = \delta_{\alpha\beta}\frac{2k_BT}{N\zeta}t$$

The self-diffusion constant of the center of mass is defined by,

$$D_G \equiv \lim_{t\to\infty}\frac{1}{6t}\sum_{\alpha=x,y,z} < (\mathbf{X}_{0\alpha}(t) - \mathbf{X}_{0\alpha}(0))^2 >$$

Thus D_G ∝ N −1, τ_1 ∝ N_2. However, this prediction is not consistent with the experimental results, which, in θ conditions, is summarized as D_G ∝ N −1/2, τ_1 ∝ $N^{3/2}$. This failure comes from the neglect to hydrodynamic interactions.

References

[1] A.V. Finkelstein O.V. Galzitskaya. Physics Of Protein Folding. *Physics of Life Reviews 1 (2004) 23–56*

[2] Liu Hong. The Statistical Models for Globular Protein Folding in Water Solution. April, 2009

[3] Eugene Shakhnovich. Protein Folding Thermodynamics and Dynamics: Where Physics, Chemistry and Biology Meet. *Chem Rev. 2006 May ; 106(5): 1559–1588*

[4] Pablo Echenique. Introduction to protein folding for physicists. February 1, 2008

[5] Pietro Faccioli. Investigating Biological Matter with Theoretical Nuclear Physics Methods. *arXiv:1108.5074v1 25 Aug 2011*

[6]http://www.csb.pitt.edu/archive/Education/Past_Courses/BioInf2052Lectures/lecture11.pdf. Classical kinetic modeling of protein folding/unfolding

[7] Stefan Wallin and Eugene I Shakhnovich. Understanding ensemble protein folding at atomic detail. *2008 J. Phys.: Condensed. Matter 20 283101*

[8] Valerie Daggett and Alan R. Fersht. Is there a unifying mechanism for protein folding? *TRENDS in Biochemical Sciences Vol.28 No.1 January 2003*

[9] Nikolay Perunov and Jeremy L. England. Quantitative theory of hydrophobic effect as a driving force of protein structure. *9 January 2014 proteinscience.org*

[10] Hongxing Lei and Yong Duan. Kinetics and Thermodynamics of Protein Folding. *http://www.intechopen.com*

[11] Kerson Huang. Lectures on Statistical Physics and Protein Folding. *World Scientific.*

[12] Geoffrey C. Rollins and Ken A. Dill. General Mechanism of Two-State Protein Folding Kinetics. *Journal of the American Chemical Society. pubs.acs.org/JACS*

END.