Protein Folding Biophysics : Theory and Experiment

Deepam Gupta, Nitai P. Sahoo, Rahul, Sayantan Chatterjee, and Vijay Ganesh*

Indian Institute Of Technology, Kanpur

This manuscript was compiled on April 21, 2017

Protein folding is an open problem in the field of both theoretical and experimental biochemistry. Numerous theoretical models and experimental techniques have paved ways for considerable developments in this field, but there isn't an integrated and all-encompassing theory that accommodates all the convoluted details, that come from empirical experimentation. In the in-vitro analysis of proteins in the presence of solvents, which is the typical theme of experimentation in experimental biophysics, we also have to account for solvent effects and this demands a Kramers' like description. In this review, we present a complete and elaborate description of this problem with emphasis on biological, theoretical and experimental viewpoints. In the course of this description, we present the specific results obtained by Chung and Eaton(1), and highlight its importance and relevance in this vast discipline.

Levinthal Paradox | Spin Glasses | Transition Path | Kramers' Theory | FRET | Maximum Likelihood Method | Internal Friction

1. Biological Background

1.1 Proteins: A general introduction .

Protein are biomolecules made up of one or more chains of amino acids. From catalyzing metabolic reactions to transporting molecules from one location to another, proteins play a vital role in biological systems. At the most fundamental level, proteins differ from one another mainly in their amino acid sequence – the various permutations and combinations of the 20 naturally occurring amino acids give rise to a vast repertoire. It has been estimated that yeast cells contain around 50 million proteins, while in human cells the count is in the order of billions (2). The instructions for the synthesis of this huge variety of proteins are embedded in every organism's genetic code.

Synthesis of a polypeptide occurs in multiple steps, starting from the deoxyribonucleic acid (DNA) sequence of our genome. The template DNA is transcribed into complementary ribonucleic acid (RNA) sequences by specialized enzymes. These newly synthesized mRNA (messenger RNA) molecules are then utilized as template for nascent polypeptide synthesis by molecular machinery called ribosomes. Peptide bonds are formed between the growing polypeptide chain and the arriving amino acid, catalyzed by the ribosomal scaffold. This results in the formation of a long chain of repeating amino acid units. This flow of genetic information (DNA $\rightarrow$ RNA $\rightarrow$ Protein), is famously called the Central Dogma of Biology. (2)

However, proteins do not exist in biological systems simply as long polypeptide chains. They possess multiple levels of structural organization, and it is this hierarchy of structure that imparts them with a biological function. (2)

- **Primary (1°)** The basic amino acid sequence is often referred to as the primary structure of a protein. The major structural feature of this structure is the planar peptide bonds, between the carboxyl terminal of one amino acid and the amino terminal of the next. The rigidity of the structure is maintained by these bonds. This is also often referred to as the unfolded form of a protein.

- **Secondary (2°)** The amino acid chains are arranged in specific structural motifs, called the secondary structure of a protein. Some examples are α-helix, β-sheet, turns, etc.

- **Tertiary (3°)** The further organization of secondary structural motifs into more complicated geometries constitutes the tertiary structure of a protein. The orientation of specific sections of the proteins into functionally important domains, for instance, is an example of the commonly tertiary level modifications. Other examples could include the occurrence of globules in many biologically relevant proteins.

- **Quaternary (4°)** Multiple peptide sequences, that have by now organized themselves into well-defined tertiary structures, often come together and form superstructures to form a biologically active protein moiety. This level of organization, seen in a lot of important proteins even in the human body, is called the quaternary structure.

The structural features of proteins have been studied extensively. Indeed, the full crystal structures of many proteins have been solved, to atomic resolution. There are a variety of experimental techniques that have been employed to investigate and characterize proteins.

- **X-ray Crystallography:** Proteins, like many other molecules that have a well-defined structure, can be crystallized. This technique is widely used to precisely solve the native structure of proteins to high resolution(3). For many proteins, the crystal structures of other molecules like DNA, RNA, small molecule inhibitors, other peptides, etc. in complex with the protein of interest have also been published(see fig.1).

- **NMR:** Nuclear Magnetic Resonance spectroscopy is widely used to characterize and identify secondary and tertiary structural features of proteins(4).

- **Circular Dichroism:** Circular Dichroism spectroscopy is mostly used(6) to identify and characterize secondary structures, like α-helix, etc.

S.C worked on the Biological part, V.G and N.P.S worked on the theory, R worked on the data analysis, D.G worked on the experiments.

The authors declare no conflict of interest.

*To whom correspondence should be addressed. E-mail: vijaygs@iitk.ac.in

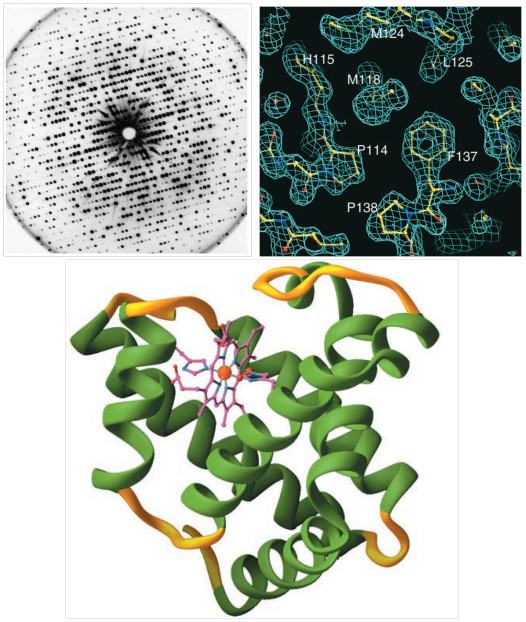

Fig. 1. X-Ray Diffraction Image, Electron density map, Computer generated cartoon image for a model protein. Image taken from ref. (5)

- **Edman Degradation:** This is a peptide sequencing technique(7). This is used, in conjunction with other powerful methods like Mass Spectrometry, to determine the primary structure of a protein, namely the amino acid sequence it is constituted of.

But why does a chain of amino acids go take the trouble to organize itself into higher order structures? While the fact that numerous interactions – hydrogen bonding, disulfide linkages, hydrophobic interactions, ionic interactions, π- stacking, Van der Waals' forces, etc – stabilize such structures, the driving factor is that for a protein, the biological activity is dependent on its three - dimensional structure. In essence, a protein devoid of secondary and subsequent higher order structural features would be useless in its biological environment, even with the correct amino acid sequence.

How an unfolded (or denatured) amino acid sequence achieves its final, biologically active, three - dimensional folded (or native) form is a question that is to be answered by investigating the process of protein folding.

1.2 Protein folding – towards native structure.

The problem of protein folding was aptly stated in the 1960s by Cyrus Levinthal(8). He conceptualized a thought experiment, now famously known as the Levinthal's Paradox which led to great insights into the process and dynamics of how proteins achieve their ultimate, biologically active, native form.

Considering a linear polypeptide sequence that is 100 residues long (which folds into the unique native structure), and assuming that each amino acid can attain 3 conformational states, it is implied that there are 3^{100} structures possible which are conformationally distinct from each other – of which only one structure is correct. In case of sequential sampling, considering just nanoseconds of time to review each conformation, the time required to achieve the native structure would exceed the age of the universe! However, most proteins effectively fold into their native structure from their unfolded amino acid sequence in biological environments in microseconds, and they are much longer than 100 residues with many more degrees of freedom. This apparent inconsistency is widely known and cited across scientific literature as the above mentioned Levinthal's Paradox.

This paradox is quite easily resolved by rejecting the assumption of sequential sampling of all the possible conformations. Levinthal himself suggested that "... protein folding is sped up and guided by the rapid formation of local interactions which then determine the further folding of the peptide; this suggests local amino acid sequences which form stable interactions and serve as nucleation points in the folding process..."(9)

The implication of this statement is the tremendous foresight that instead of just probabilistic processes that require sampling over all the possible conformations, the kinetics of protein folding may be elucidated using the idea that proteins utilize specific pathways to reach their native structure from their unfolded form. This, in essence, implies the possibility of observing and characterizing intermediates in the process of protein folding – and such structures have indeed been found in multiple studies(10).Interestingly, it has been discovered that certain special proteins called chaperones assist in the process of protein folding and unfolding(11). The importance of the presence of these proteins is underlined by the fact that in the highly fluxional biological environments where proteins undergo folding, very often the polypeptide chain veers off course and misfolds. If it were not for these molecules that unfold and refold these off-track proteins, every minor misfolding event would render the polypeptide moiety functionally useless. Such mechanistic models of protein folding also rely upon the nucleation of amino acid chains into precursor secondary structures. Experiments have established that indeed small regions of the polypeptide chain do undergo structural organization into defined nucleation centers, which then go on to further drive the process of folding(12).

Considerable research has been undertaken towards understanding the thermodynamics of protein folding as well. An initial foray into this domain was presented by Christian Anfinsen in 1961 (Nobel Prize, 1972)(13). He proposed the famous thermodynamic hypothesis of protein folding – that it was sufficient to know the amino acid sequence of a protein to deduce its native three-dimensional structure. An important implication of such a hypothesis was understanding that the folded protein would be in thermodynamic equilibrium with the biological environment. The conclusion arrived at was conceptually very elegant – all one had to do, apparently, to arrive at the native structure of a protein was to sum all the interatomic interactions in every possible conformation, and the conformation with the lowest internal energy would be the native form! While this hypothesis has undergone further refinement, it is certainly not an easy task to carry out such calculations. However, there have been great strides in such research guided by Anfinsen's hypothesis(14).

The process of protein folding is guided by multiple varieties of local as well as non-local interactions. There are many non-covalent interactions (as mentioned earlier) that occur to stabilize a protein structure, but the important point is that they need not be spatially close to each other. There are many non-covalent interactions that are long range, and

have been characterized by experiments(15). Many types of long-range interactions have often been credited by theoretical models as well as experimental results with providing significant stabilization to peptide motifs. Some of the most well characterized non-local interactions have been electrostatic interactions, especially ion-dipole and dipole-dipole interactions between amino acid residues that are far from each other in the sequence(16).

With the availability of accurate instrumentation and increased computing power, it is now possible to investigate the dynamics of protein folding itself using various experimental techniques and simulations. Various studies have also been published, establishing the concurrence of theoretical/computational results with experimental data(17).

Probing the dynamics of protein folding is, thus, a very challenging area of research in the biological sciences as well as in the allied fields. The paper under consideration in this discussion is indeed an endeavor towards this goal.In the next section we discuss the theoretical aspects of the protein folding problem.

2. Theory of Protein folding Biophysics

2.1 Protein folding as a phase transition problem.

Contemporary experiments in structural biology have affirmed the existence of a hierarchy of forces operating between the amino acid residues over various length scales, which was briefly discussed in the previous section. The existence of local interaction forces operating between pairs of amino acid residues, which decisively affects the final folded structure, is reminiscent of spin glasses, which is a generalisation of the Ising model for ferromagnetism(18) where pairwise interactions are specified. The model is appropriate, as minute changes in structural parameters such as the orientation of the amino acid residue could quite drastically affect the final folded structure. Such a conception in terms of spin glass had been formulated by PG Wolynes and Bryngelson(19), who use the random energy model in the same spirit as used by Derrida, to analyze spin glasses(20).

Typically, it is observed that small proteins transform discontinuously from the unfolded state to folded state which is suggestive that a phase transition is operative. Since this system of amino acid chain has finite degrees of freedom, phase transition is obtained in the infinite size limit where singularities show up in the free energy function. In large proteins, 'molten globule' intermediates, which are thermodynamically distinct from native and denatured states, are also observed(21). In this regard, the spin glass model is quite relevant, as they show that it allows for the existence of a 'glassy' state below a characteristic temperature.

Apart from the structural aspects associated with the native state optimization process, the kinetics observed in protein folding are quite bewildering. The protein folding process, which occurs in under a second, neither proceeds through random sampling over the ensemble of configurations available, which was stated in the previous section as 'Levinthal Paradox', nor as a search for a potential minima in a smooth potential energy surface, as that cannot be consistent with the general multiexponential kinetics observed. A detailed analysis of folding kinetics is provided in the next section.

The spin glass Hamiltonian for the amino acid chain is:

$$E = -\sum_i \varepsilon_i(\alpha_i) - \sum_i J_{i,j}(\alpha_i, \alpha_{i+j}) - \sum_{i,j} K_{i,j}(\alpha_i, \alpha_j, r_i, r_j) \quad [1]$$

The three energy terms in the above expression, account for the hierarchy of interactions in the chain, the first term being associated with the primary structure, the second with the secondary structure and the third with the tertiary structure. Each amino acid residue has a native state, in which it exists when the protein is in the native folded state, and 'v' other states and it is established that 'v' is of the order of 10 residues.

A set of assumptions are made after the Hamiltonian is set up: The underlying evolutionary biological process, would not let small perturbations in the final folded structure affect the native protein drastically. With the analogy from solid state physics already in place, the perturbations in the positions of the amino acid residues in any given state is termed as 'frustration' much similar to the geometrical frustrations(22), which alter the preciseness of the ground state energy in crystals. The native state should hence be the structure with minimal frustration and such an assumption is technically named as the 'Principle of minimum frustration'. This principle is very powerful in that it singly eliminates the Levinthal paradox by establishing a bias towards the native state.

With the number of amino acids in a chain being of the order of 100 or more, it is not easy to handle the complicated Hamiltonian and assume Boltzmann statistics for the microstates in the usual sense. A replacement of this complex Hamiltonian with a stochastic Hamiltonian, similar to the idea of Wigner random matrices(23) is suggested and thus the 'Random energy model' assumes a central role in describing the underlying crucial details of the Hamiltonian.

A probability distribution in terms of the number of native amino residues and the energy is set up and the energies of individual protein conformations are assumed to be uncorrelated.

$$P(E_1, E_2 ... E_n, N_0) = \prod_{i=1}^{n} (P_i, N_0) \quad [2]$$

The native primary, secondary and tertiary interactions are denoted by -ε, -J, and -K respectively and the non-native interactions are distributed with mean -$\overline{K}$, -$\overline{J}$ and -$\overline{\varepsilon}$ and standard deviation $\triangle K$, $\triangle J$ and $\triangle \varepsilon$. The number of residues neighbouring each residue is denoted by z. In terms of these parameters, the mean total energy expression is:

$$\overline{E}(N_0) = -N_0\varepsilon_0 - \left(\frac{N_o^2}{N}\right) L - (N - N_0)\overline{\varepsilon} - \left(N - \frac{N_0^2}{N}\right) \overline{L} \quad [3]$$

where $L = J + zK$.

The average density of states expression in the thermodynamic limit, as obtained in terms of the fraction of native residues, from the stochastic Hamiltonian is :

$$log\,\langle n(E)\rangle = N \max_{0\leq p\leq n}\left\{-\rho\log\rho-(1-\rho)\log(1-\rho)\log\left(\frac{1-\rho}{\nu}\right)-\frac{[\varepsilon+\overline{\varepsilon}+\overline{L}+(\rho_0-\overline{\varepsilon})\rho+(L-\overline{L})\rho^2]^2}{2[\Delta\varepsilon^2+\Delta L^2-\Delta\varepsilon^2+\Delta L^2-\Delta\varepsilon^2\rho-\Delta L^2\rho^2]}\right\}$$

This quantity can be understood to be the entropy of the chain, dependent on the energy and the fraction of native residues. The expressions for the free energy per amino acid residue(f) and entropy per residue(s) (last two terms) are obtained as:

$$f=-\overline{\varepsilon}-\overline{L}-\frac{\Delta\varepsilon^2+\Delta L^2}{2T}-\left(\varepsilon-\overline{\varepsilon}-\frac{\Delta\varepsilon^2}{2T}\right)\rho-\left(L-\overline{L}-\frac{\Delta L^2}{2T}\right)\rho^2+T\rho\log\rho+T\,(1-\rho)\log\left(\frac{1-\rho}{\nu}\right) \quad [4]$$

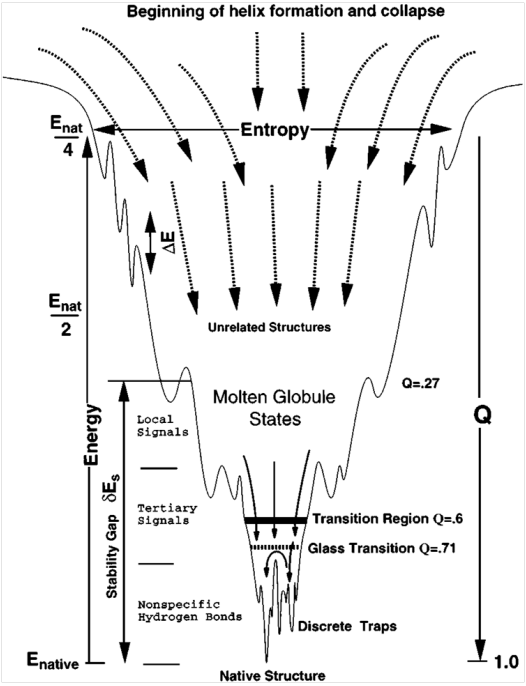

Fig. 2. The rugged and funneled energy landscape, with a lot of local minima. Molten globule states and glass transition point are marked. The native state is marked by least energy and least entropy. The fluctuations in the energy, in accordance with the random energy model manifest themselves in the ruggedness of the landscape. Image taken from ref. (24)

It is easy to observe that at a characteristic temperature, hereby termed as 'Glass transition temperature(T_g) ', the entropy goes to zero and hence the free energy attains the maximum value. The phase below this glass transition temperature is the glassy phase, which manifests as misfolded states in folding experiments as was mentioned before.

$$T_0=\left\{\frac{\Delta\varepsilon^2+\Delta L^2-\Delta\varepsilon^2\rho-\Delta L^2\rho^2}{2\left[-\rho\log\rho-(1-\rho)\log\left(\frac{1-\rho}{\nu}\right)\right]}\right\}^{\frac{1}{2}} \quad [5]$$

In condensed matter physics, phase transitions are marked by a discontinuity either in the value of free energy or in the value of some order derivative of the free energy with respect to the 'order parameter'(25). The existence of dynamic

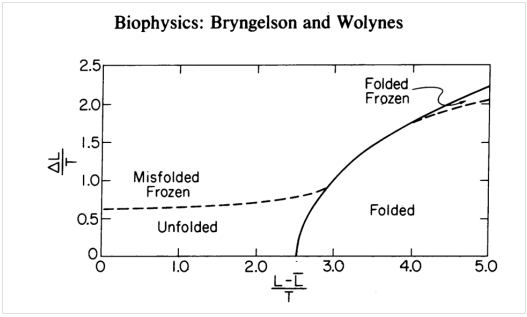

Fig. 3. In this figure, the fluctuations in the native tertiary and secondary structure interaction energies, accounted by the fluctuations in the collective coordinate L, is plotted against the difference between the energies of the particular state and the native state. As the difference increases, the structure becomes more 'native-like' and hence native folded phase is observed at higher values of the difference. The curves, both dotted and solid, denote phase transitions where a singularity sets up in some nth order derivative of the free energy. Figure taken from ref.(19)

equilibrium between phases is consistent with the observation of equilibrium between the folded and unfolded states. The phase transition from folded to unfolded state is a first order phase transition, as there is a singularity in some n^{th} order derivative with respect to the order parameter at a specific glass transition point.The phase diagram along with the phase transitions is provided in figure 3 and 4.

This order parameter is crucial to the characterization of a condensed phase, and in the case of proteins, the choice of the order parameter - the fraction of native state amino acid residues is quite powerful and such a conception proves to be useful in understanding the dynamics of the folding process too. The analogue drawn here between the peptide chain and a spin-glass, apart from serving as an explanation to the numerous mis-folded states observed in the folding process, also explains the general kinetics observed in folding experiments.

The random energy model assumption and the stochastic nature of the Hamiltonian is marked by pronounced fluctuations(26), which result in the roughness of the so called 'Energy Landscape'. The energy landscape is essentially the free energy of the configurations as a function of a global parameter, in the case, the fraction of native residues. The landscape, as can be observed in figure 2, has a lot of local minima, which hinders the 'roll-over' to the native state. The

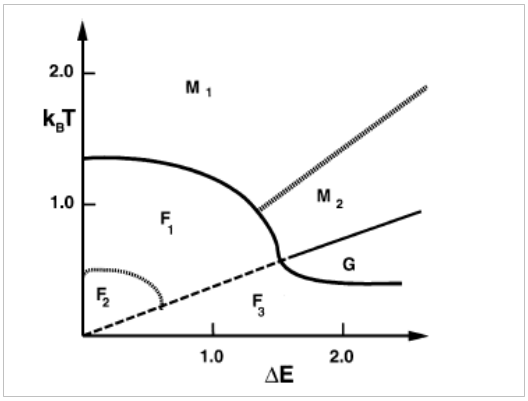

Fig. 4. Typical phase diagram of a Random heteropolymer. The dashed line corresponds to the glass transition. Solid curves separate static phases and dotted lines refer to metastable unfolded states, which transiently reside in the local minima. M1, M2, M3 correspond to the molten globule states, F1, F2 and F3 correspond to the folded states. G refers to the glassy phase. Plot taken from ref.(27)

kinetics of folding can be derived from this landscape and typically is quite complicated. In the next section, using the same random energy model, we describe the complicated kinetics associated with the folding process.

2.2 Dynamics of protein folding.

As was briefly stated in the previous section, the random energy model and the associated energy landscape predicts a funneled energy landscape with a lot of local free energy minima, which accounts for the general multi-exponential folding times observed(28). In this section, we obtain explicit expressions for the mean first passage times from these potential energy minima and aptly describe the dynamics of the folding process.

In a work which is quite relevant to the folding process we are concerned with, Zwanzig analysed the problem of diffusion in a rough potential around a minima(29). He formulated this process as a diffusion in an effective one dimensional one dimensional potential, with the averaged effects of the fluctuation and a spatially and temperature dependent diffusion coefficient and derived expressions for mean first passage times. The corresponding expressions are:

$$D^* = D\ exp\left[-(\varepsilon/k_BT)^2\right] \qquad [6]$$

$$U^*(x) = U_0(x) - \psi^-(x)/\beta \qquad [7]$$

Where U_0(x) is the smooth overall potential and U_1(x) is the low amplitude fluctuation.

$$\left\langle e^{-\beta U_1(z)} \right\rangle \ = \ e^{\psi^-(z))} \qquad [8]$$

The expression for mean first passage time:

$$\langle t, x\rangle \cong \int_{x_0}^{x} dy\ e^{\beta U_0(y)+\psi^+(y)}\ (1/D) \int_a^y dz\ e^{-\beta U_0(z)+\psi^-(z)} \qquad [9]$$

Wolynes and Bryngelson(30) generalised this formulation for a generic multi-dimensional case, which is more relevant to complex chemical reactions, such as that of protein folding. Their conception started with the same old paradigm of replacing the complex multi-dimensional Hamiltonian with a stochastic Hamiltonian which had the same characteristics as the original one.

The stochastic Hamiltonian allows for the existence of numerous local minima and explicit residence time and escape time expressions could be derived.

The model for understanding the folding dynamics is as follows: As in the previous case of structure prediction and characterization, the fraction of 'native' amino acid residues ($\rho = N_0$/N), which was assumed to be the order parameter in the previous is the variable of interest here. In terms of this order parameter, the distribution of the energies is:

$$g(E,\rho)\ dE = [2\pi\Delta E(\rho)^2]^{-1/2}\ exp\left\{-\frac{[E-E(\rho)]^2}{2\Delta E(\rho)^2}\right\} dE \qquad [10]$$

Where the terms E bar and delta E are the same as defined in the previous section.

A transition from one protein configuration to another is described as a unit change in the number of native amino acid residues. That is, the configuration with ρN native residues can transition to one with ρN-1 and ρN+1 native residues. The transition rates from one state to another is assumed to be described by a metropolis dynamics:

$$\begin{aligned} R &= R_0\ exp[-(E_B - E_A)/T] \qquad for\ E_B > E_A \\ &= R_0 \qquad for\ E_B < E_A \end{aligned} \qquad [11]$$

Where E_A is the energy of the present configuration and E_B is the energy of any configuration to which the current 'state' is connected to (Nν of them in total).

The local minima are rightly assumed to be the states, which has the least energy among the other states it is connected to. Distribution of the average escape rates is derived from transition rates distribution for such local minima:

$$P_{LM}(R,\rho) = \left(\frac{1}{2\pi}\right)^{(1/2)} \left(\frac{1}{R_0}\right) \left(\frac{T}{\Delta E(\rho)}\right) exp\left\{-\frac{T^2 \log^2\left[\frac{R}{R_{sep}(\rho)}\right]}{2\Delta E(\rho)^2}\right\} \qquad [12]$$

where

$$R_{sep}(\rho) \equiv \overline{R}_0 N\nu\ \exp[-(\Delta E(\rho)^2)/2T^2] \qquad [13]$$

Two critical temperatures are observed, one of which is the same glass transition temperature deduced from the spin glass model for describing phase transition.

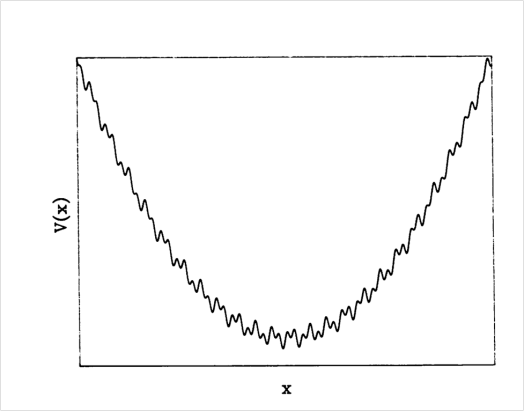

Fig. 5. The form of rough potential Zwanzig considered in his paper on 'Diffusion in a rough potential' (29). The potential is parabolic with a small amplitude oscillatory potential superimposed.

$$T_g(\rho) = \frac{\Delta E(\rho)}{[2S^*(\rho)]^{(1/2)}} \qquad [14]$$

Where $S^*(\rho)$ is given by :

$$S^*(\rho) = \log \Omega\ (N\rho) = N[-\rho \log \rho - (1-\rho)\ \log((1-\rho)/\nu)] \qquad [15]$$

Which is exactly the same as given by eq. 4.

Below this temperature, it is shown that the dynamics are very slow and comparable to one predicted by a pure random energy model. It is in this limit that the folding time approaches the Levinthal limit. Thus, the glass transition temperature(and the hence the existence of a glassy phase), obtained from contrasting the transition dynamics is same as the temperature which marks the equilibrium glass transition.

The transition from one local minima to another occurs at arbitrary times, and hence a continuous time random walk model is appropriate to characterize this 'jumping' process. The waiting time distribution in this case is given by:

$$\Psi(t,\rho) = \int_0^\infty dR\ P(R,\rho) R \exp(-Rt) \qquad [16]$$

The expression for free energy is similar to that in the previous section(eq. 5). The master equation is set up and the whole derivation is done in the 's' space after evaluating the laplace transform of the master equation:

$$\Psi(t,\rho) = \lambda(\rho)\ \Psi(t,\rho) + [1-\lambda(\rho)] \int_0^t dt' \Psi_0(t-t',\rho)\ \Psi(t',\rho) \qquad [17]$$

Where

$$\lambda(\rho) = 1/\nu + (1-1/\nu)\rho \qquad [18]$$

A generic Fokker Planck equation is derived for this specific case:

$$\frac{\partial}{\partial t} G(\rho,t) = \frac{\partial}{\partial \rho} \left\{ D(\rho) \left[G(\rho,t) \frac{\partial}{\partial \rho} \beta F(\rho) + \frac{\partial}{\partial \rho} G(\rho,t) \right] \right\} \qquad [19]$$

From here, the evaluation of mean first passage times is straightforward and is carried out in similar spirit as in Kramers' classic paper(31) and the final expression is:

$$\bar{t} = \int_{\rho_i}^{\rho_f} d\rho \int_0^\rho d\rho' \frac{\exp[\beta F(\rho) - \beta F(\rho')]}{D(\rho,0)} \qquad [20]$$

With clever manipulation, this expression can be modified to:

$$\bar{t} = \frac{1}{D_0} \int_{\rho_i}^{\rho_f} d\rho \int_0^\rho d\rho' \exp[\beta \overline{F}(\rho) - \beta F(\rho')] \qquad [21]$$

where

$$X(\rho) = \log \left[\frac{D(\rho,0)}{D_0} \right] \overline{F}(\rho) = F(\rho) - TX(\rho) \qquad [22]$$

If a double well function is assumed for $\overline{F}\ (\rho)$, then the mean first passage time is:

$$\bar{t} \approx \left(\frac{2\pi}{\beta} \right) \frac{1}{D_0 \omega_0 \overline{\omega_t}} \exp\{\beta[\overline{F}(\overline{\rho_t}) - F(\rho_b)]\} \qquad [23]$$

Where ω_b is the harmonic oscillator frequency near the 'unfolded' well and wt is the frequency associated with the barrier. This mean first passage time can be considered as the 'folding time' for the protein and this expression shall be used late.

The average lifetime of a configuration with Nρ native residues is calculated as:

$$\tau(\rho) \equiv \frac{1}{\lambda(\rho)} \left\langle \frac{1}{R} \right\rangle (\rho) \qquad [24]$$

For the glassy phase, the time required for folding becomes:

$$t \sim \sum_{N_\rho=1}^{N} \tau(\rho) \sim \frac{1}{R_0 N \nu} \sum_{N\rho=1}^{N} \Omega(N\rho) \sim \frac{(\nu+1)^N}{R_0 \cap N\nu} \qquad [25]$$

Which is of the order of the time required for random sampling. From this expression, the transition rates from glassy phase is obtained and found to be close to the Levinthal limit of folding time. On the other limit, when the temperature is higher than the glass transition temperature, they reproduce Zwanzig's expression for the effective diffusion constant, which exhibits super Arrhenius behaviour(Eq. 7).

In this derivation, we have assumed that $\overline{F}_(\rho)$ has a double well form. $\overline{F}$ can be interpreted to be the free energy of the ensemble of protein states with the same fraction of native residue. In the next section, we claim that a one-dimensional free energy surface suffices to completely describe the transition from an ensemble of unfolded states to an ensemble of folded states, separated by an ensemble of transition states. We then proceed to show how to construct such a surface using simulations and empirical experimental data.

2.3 Free energy surface for protein folding.

The free energy surface is perhaps the simplest way of global representation of the protein folding problem as a function of one collective coordinate, or a small set of coordinates. As a crudest form of explanation, we can claim that there exists a barrier in the free energy surface between the unfolded states and the folded states, by the following argument: The

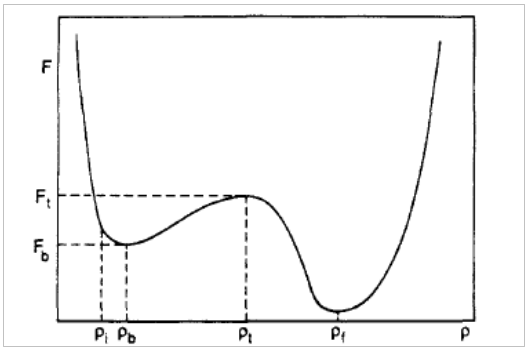

Fig. 6. One dimensional free energy surface assumed for the function $\overline{F}(\rho)$. The barrier frequency and the 'reactant well' frequency are depicted. Figure taken from ref. (30)

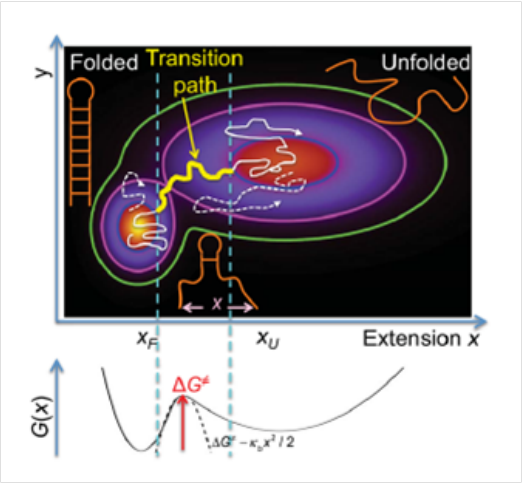

Fig. 7. A sample trajectory in the transition path region is given. The corresponding free energy diagram is given just below. Image is taken from ref.(35)

unfolded structure, although being high in energy, as inferred from the energy landscape, has a lot of possible micro-states in which it can exist and hence has a large entropic contribution and hence has a net low free energy. On the other hand, the native structure has one unique configuration (or a relatively less number of possible configurations) but has a very low energy. Hence the free energy is again minimal. Since there is a trade-off between entropic contribution and total energy contribution, it is reasonable to claim that there is a barrier in the free energy surface between the ensemble of folded and unfolded states.

The free energy surface can be characterized by one global parameter, for instance the fraction of native amino acid residue, which was used in the sections before, or a set of global parameters(32). In such a method, the energy landscape is stratified and every possible configuration in a strata is represented by a particular value of global parameter. It has been proven by experimental and computational studies that a one-dimensional free energy surface is sufficient to describe the folding dynamics completely and quite accurately(32).

The typical workhorse in such simulation studies has been the 27-mer and 36-mer lattice model(33). The global parameter is chosen the number or fraction of native amino acid (Q/ρ) residues in such a model. At every possible value of Q, the total energy of the polypeptide chain and the number of possible configurations-micro-states and hence the entropy, allowed at that specific value of Q is calculated and the free energy is obtained. It was shown by Socci et al(32) that if the glass transition happens after the 'transition state' in free energy surface, the analysis of Wolynes and Bryngelson is appropriate. This is due to the fact that once glass transition happens, the folding time approaches the Levinthal limit and the calculations lose relevance.

The folding time depends not only on the height of free energy barrier but also on the probably spatially dependent diffusion coefficient. The diffusion coefficient was obtained by calculating the mean square fluctuations of the configuration variable 'Q' and correlation time:

$$C_Q(\Delta) = \frac{\langle Q(t)Q(t+\Delta)\rangle - \langle Q(t)\rangle^2}{\langle Q^2(t)\rangle - \langle Q(t)\rangle^2} \quad [26]$$

$$D = \Delta Q^2(T)/\tau_{corr}(T) \quad [27]$$

The analytical expression as was mentioned before(29), for the diffusion constant is

$$D(T,n) = D_0 \exp[-\beta^2 \Delta E^2(n)] \quad [28]$$

for high temperatures. For intermediate temperatures, the modified expression(30) is:

$$D(T,n) = D_0 \exp\{-S^*(n) + [\beta_g(n) - \beta]^2 \Delta E^2(n)\} \quad [29]$$

The mean passage time was obtained as a function of temperature and large values of it were observed at temperatures less than the glass transition temperature, proving the predictions of Wolynes and Bryngelson. The T_f/T_g ratio gives a measure of how 'good' a protein folds and has emphasized in literature(34).

Apart from these kinetic calculations, thermodynamic quantities were calculated too: Histograms of (Q,E) points were plotted and the corresponding density of states n(Q,E) and hence the free energy and other thermodynamic quantities were evaluated.

Now, with the free energy surface computed and the diffusion coefficient calculated, the validity of the first mean passage times expression derived by Wolynes and Bryngelson could be checked. The analytical expression proved to be a very precise estimate of the folding times.

The one dimensional free energy surface is quite rough as can be observed. It is then quite tempting to assign the maxima in the free energy surface to be the 'transition state', which is however not quite correct as it was not 'recrossing free', which is the signature of transition states. It is hence relevant to consider a 'transition region' instead of a single point in the free energy surface. Such a conception calls for a rigorous theoretical understanding, which shall form the crux of our next section.

2.4 Transition path formalism of transition state theory.

The transition path formalism, due to Hummer(36) identifies a specific region in the potential energy surface, which connects non-recrossing trajectories exiting from 'Reactants region' and entering 'Products region' and contains no points from either of the regions, as a transition path ensemble. Individual transition paths are the trajectories which connect

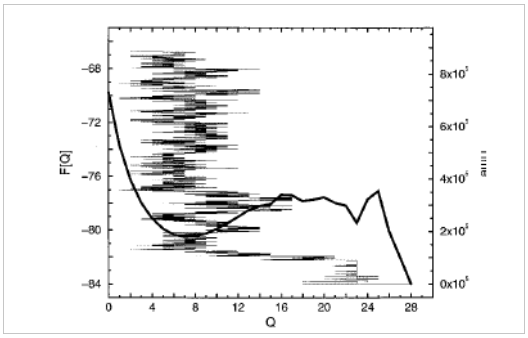

Fig. 8. The free energy diagram obtained from thermodynamic calculations, plotted with outputs from folding trajectory simulation. The transition region can be observed to be rough. Figure taken from ref.(32)

these points.Transition states are defined as 'points in this region where probability that the trajectories passing through them are reactive is maximum'. The formalism is generic in that the transition paths can be appropriately set up in both phase space and configuration space, and also regardless of whether the underlying is deterministic or stochastic.

A conditional probability density (p(x/TP)) is set up in the transition path ensemble, where x denotes a position in phase space or configuration space.Given that the system is in x, the probability for being on a transition path can be evaluated as:

$$p(TP|x) = \frac{p(x|TP)p(TP)}{p_{eq}(x)} \quad [30]$$

Where p(TP) is the fraction of time spent in the reaction path and p_{eq}(x) is the equilibrium Boltzmann distribution. In the case of langevin dynamics in configuration space, the exact expressions can be obtained, while for dynamics in phase space, a projection onto the reaction coordinate is to be used. The exact expressions for the conditional probability density is obtained for the case of high friction limit of Langevin equation.

The conditional probability density (p(x/TP) can be evaluated for the high friction limit of Langevin equation, from the appropriate solution of the corresponding Smoluchowski equation, assuming that the particle is 'absorbed' at the boundaries(i.e p(x/TP) = 0).

$$\frac{\partial G}{\partial t} = D\frac{\partial}{\partial x}\left\{e^{-\beta V(x)}\frac{\partial}{\partial x}[e^{\beta V(x)}G]\right\} \quad [31]$$

Where V(x) is the potential energy function and D is the diffusion coefficient. This is then related to the 'splitting probability' (ϕ_A(x) or ϕ_B(x)) which is defined to be the probability that a particle in the transition path reaches the product region/ reactant region first. The probability distribution for the time spent in the transition region is obtained, from which the average transition path time is evaluated as:

$$\langle t_{TP}\rangle = \frac{\int_{x_0}^{x_1} e^{-\beta V(x)}\Phi_A(x)\Phi_B(x)dx\int_{x_0}^{x_1} e^{\beta V(x')}dx'}{D} \quad [32]$$

An analytical expression for the average transition path time for various potentials was derived by a slightly different conception by Kim and Netz(37). Forward and Backward Fokker-Planck equation(38) was used to derive analytical expressions for mean first passage time, splitting probabilities and hence by a limiting procedure, the mean transition path time.

The Forward Fokker Planck Operator is:

$$L(x) = \partial_x D(x)e^{-F(x)}\partial_x e^{F(x)} \quad [33]$$

Where D(x) is the spatially dependent diffusion coefficient and F(x) is the free energy. The corresponding solution, with the same boundary conditions as assumed by Hummer and its time evolution in terms of the adjoint operator gives:

$$\partial_t G(x,t|x_0) = L^\dagger(x_0)G(x,t|x_0) \quad [34]$$

The outgoing 'diffusion current' from the transition region into the reactant or product region is a measure of the first passage rate. The average diffusion current can be obtained from the Smoluchowski diffusion equation as

$$\left\langle j_{A/B}(x,t|x_0)\right\rangle = \int_0^\infty dt j_{A/B}(x,t|x_0) = j(x_{A/B}|x_0) \quad [35]$$

And hence the total current leaving the transition is j_A + j_B. The average current leaving the region is given by:

$$j(x_{A/B}|x_0) = \int_0^\infty dt\; j(x_{A/B},t|x_0) \quad [36]$$

In the infinite time limit, the integral of diffusion current would would be equal to unity as all the 'flux' contained inside the transition region would go either into the product or reaction region at infinite time. This gives the corresponding splitting probabilities ϕ_A and ϕ_B. After normalization, the mean passage time is given as:

$$\tau^{FP}(x_{A/B}|x_0) = \frac{j(x_{A/B}|x_0)}{\phi_{A/B}(x_0)} \quad [37]$$

Explicit expression for the splitting probabilities can be obtained by calculating survival probabilities and expressing the current as a time derivative of the survival probabilities.

$$S(x_0,t) = \int_{x_A}^{x_B} dx G(x,t|x_0) \quad [38]$$

$$\partial_t j(x_{A/B},t|x_0) = L^\dagger(x_0)j(x_{A/B},t|x_0) \quad (40) \quad [39]$$

Thus the equations for the splitting probabilities yield:

$$e^{F(x_0)}\frac{\partial}{\partial x_0}D(x_0)e^{-F(x_0)}\frac{\partial}{\partial x_0}\phi_B(x_0) = 0 \quad [40]$$

$$\phi_A(x_0) = 1-\phi_B(x_0) = \frac{1}{C}\int_{x_0}^{x_B} dx\frac{e^{F(x)}}{D(x)} \quad [41]$$

$$C = \int_{x_A}^{x_B} dx\frac{e^{F(x)}}{D(x)} \quad [42]$$

So the final form of expression for mean first passage time to reach B is:

$$\tau^{FP}(x_B|x_0) = C\frac{\phi_A(x_0)}{\phi_B(x_0)}\int_{x_A}^{x_0} dx e^{-F(x)}\phi_B^2(x)\cap + C\int_{x_0}^{x_B} dx e^{-F(x)}\phi_A(x)\phi_B(x) \quad [43]$$

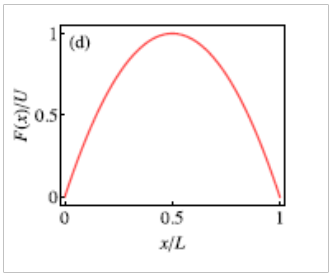

Fig. 9. The inverted harmonic oscillator potential, used to approximate the barrier region. Figure taken from Ref. (37)

In the limit of x_0 tending to x_A, the transition path time is obtained, exactly as eq. (33).

$$\tau^{TP}(x_B|x_A) = C\int_{x_A}^{x_B} dx e^{-F(x)}\phi_A(x)\phi_B(x) \qquad [44]$$

Kim and Netz have derived explicit expressions for a few model free energy functions, including one for an inverted harmonic oscillator of the form shown in figure.In the limit of a large barrier height, the transition path time for an inverted harmonic oscillator potential is obtained as:

$$\tau^{TP}(L|0) \approx \frac{\ln(2e^{\gamma}U)}{D\omega^2} \qquad [45]$$

Where ω^2 is the second derivative of the function F(x) at the barrier top. The Euler constant (gamma) comes from the infinite limit of the hypergeometric function which occurs as a solution of the integral.

Transition path time measurement is a topic of experimental interest, as it could reveal crucial details about the nature of the free energy surface of the folding protein. A lot of work has been going on in analyzing the protein folding problem using single molecule fluorescence based experiments. In the next section, we summarize the discussion on the theory of protein folding and set up the tone for the experimental aspects of the problem.

2.5 Connecting theory and experiment: The story so far....

In the last few sections, we had put forward an almost overarching discussion on the theory of protein folding. The spin-glass model, together with the random-energy assumption and the principle of minimal frustration was shown to explain the robustness of the native structure. It also explained the formation molten globules states or intermediates, observed in folding experiments as manifestations of the glassy phase and drew an analogy to the case of phase transitions in finite systems, obtained in the infinite size limit. A characteristic glass transition temperature was found to exist below which glassy dynamics can be observed.

The same paradigm of replacing the complex Hamiltonian with an all-encompassing stochastic Hamiltonian was used to analyse the dynamics associated with the folding problem. The existence of various local minima was shown and expressions for average resident time in the local minima was obtained. The average time taken for the unfolded protein to fold to its native form, through a lot of 'folding pathways' inundated with local energy minima, is expressed in terms of mean first passage time a-la Kramers using the theory of continuous time random walks. It was also shown that in the glassy phase, the folding time could approach the Levinthal limit. The appropriate expression for 'diffusion coefficient' and its temperature dependence were obtained. The final folding time expression for a double well shaped free energy surface was obtained, which was analogous to Kramers' expression.

The claim that a one-dimensional free energy surface parametrized by a single collective coordinate can completely describe the dynamics is supported by lattice simulations, which were accordingly explained. The free energy surface was observed to be quite rough and the free energy maxima did not correspond to the transition state bottleneck, in the usual sense. It was suggested that a transition 'region' is used instead of the usual conception of saddle point transition state. Such an idea called for theoretical attention and the transition path formalism was described. The expression for the average transition path time for an inverted harmonic oscillator potential was obtained.

Given all this, we can now discuss those aspects of theory that can be subjected to direct experimentation. The ratio of average transition path time to average folding time would give an accurate estimate of the free energy barrier height, given all other parameters. Thus, the measurement of transition path time and folding time gives an accurate estimate of free energy barrier height. Precisely, these two times are measured in single molecule fluorescence experiments. The ratio of barrier frequencies are obtained from molecular dynamics simulation. The various intricacies associated with the experimental procedures are explained in the upcoming sections.

3. Protein folding problem in experimental biophysics:

Our current experimental understanding of the dynamics of the folding process is quite limited as most experimental tools used to probe dynamics are indirect. Most of the experiments provide specific information about stable intermediates and miss out most of the dynamical aspects of the folding process. Several results are conjectured based on empirical experimental data(39). An analogue of Femtosecond Transition State Spectroscopy(FTS)(40), which captures reaction progress at short intervals of time, is not available for proteins, to record a complete folding trajectory. Such a technique would essentially reveal all characteristics features of an energy landscape. Nevertheless, there has been a lot of progress in experimental biophysics and several theoretical predictions have been confirmed. In this section, we present a chronology of experimental advancements in biophysics and describe the extent to which theory has been connected to experiments.

Conventional protein dynamics studies had been based on ensemble averaged properties of the folding proteins. These are often referred to as 'Ensemble experiments' (41) and they record bulk properties of the proteins at various instances in the folding process. In perturbation-relaxation approaches(15), a population of protein molecules is perturbed by a sudden rapid change of a physical parameter that could affect the equilibrium population decisively. As the perturbation is relaxed, the molecules slowly revert back to their equilibrium distribution and the relaxation process could reveal relevant details about the dynamics. These techniques are referred to as 'Ultrafast mixing' and include temperature jump methods, laser based methods, quenched flow methods etc (15). Photochemical methods which don't disturb the protein folding

process and uses auxiliary probes, have also been deviced(15).

The first protein to be crystallised was myoglobin in the year 1958 and since then many attempts have been made since to use X-Ray Crystallography to study the dynamical aspects of protein folding by observing the intermediates. However, these attempts failed as these methods were intrinsically slow, and also required that folding intermediates are thermodynamically stable enough to be recorded. In general, at equilibrium, intermediates are not significantly populated for single domain proteins, most likely due to marginal stability of intermediates and hence such experiments were not viable.

Many standard methods were then developed to study protein folding:

- **Analysis of folding intermediates by Circular Dichroism (CD)** Various studies(42) have shown that the folding pathway is seeded by local folding of domains which are termed as 'kernels' if they are stable and 'nucleus' if they are not. By CD we can monitor the kernels.Depending on the perspective, equilibrium or kinetic (stopped flow) CD has been used, in connection with various chemical and biochemical modifications of the proteins, to search for intermediates. This method studies the kinetics of folding and has the advantage that they can be carried out in conditions where the folded state is strongly favored and intermediates are more likely to be stable. Although well-populated intermediates have been detected in kinetic folding experiments, the intermediates are transient and steps in folding are often fast and hence probing dynamics is not quite easy or reliable using this method.

- **Hydrogen-Deuterium Exchange method (HDX)** The relatively poor resolution of NMR and slow detection issue in case of X-Ray crystallography, led to the development of a new method based on H-D exchange(43). HDX is an isotopic labeling technique in which amide hydrogens in the peptide backbone become "exchanged" for deuterium upon incubation in D_2O-containing buffer. Regions of the protein that are exposed to solvent readily exchange at the earliest time of incubation, allowing HDX to probe the structure of a protein. Furthermore, local fluctuations result in transient exposure of otherwise non exchangeable amide hydrogens. Over time, these transient exposure events result in increasing deuterium incorporation at any given site, a direct measurement of conformational dynamics. When HDX is coupled with proteolytic digestion and MS, the extent of deuterium incorporation can be localized to specific peptides. Supplemented by Molecular Dynamics Simulation under related conditions,which can pinpoint individual residues and dynamic behaviors responsible for specific observations, more direct observation of the process is possible than by X-Ray crystallography.

As was mentioned before most of these methods are indirect and they estimate bulk properties of the folding proteins. The dynamic properties of the ensemble of unfolded states as the folding process progresses cannot be deciphered from these ensemble experiments. In the next section, we describe a new paradigm of experimentation termed 'Single Molecule Experiments' and its relevance in understanding the folding problem.

3.1 Single molecule Experiments: FRET.

With the optical detection and spectroscopic characterization of single molecules in condensed matter made possible by the seminal works of Moerner et al(44), there was a paradigm shift in the approach of experimental research in the condensed matter community. The power and versatility of single molecule experiments in biology was realised very soon(45) and renewed research endeavours grew in the problem of protein folding. Fluorescence techniques were found to yield the best results among the numerous single molecule techniques due to its implementational simplicity and molecular specificity. We describe the phenomenon of Forster Resonance Energy Transfer(FRET) and its usage in single molecule experiments.

FRET. The basic principle associated with the FRET experiments is the phenomenon of non-radiative energy transfer between two light sensitive chromophores, through dipole-dipole coupling. The efficiency of such a dipole coupled energy transfer is inversely proportional to the distance between the two chromophores. The expression for efficiency is given to be:

$$E(r) = \frac{1}{1 + (r/R_0)^6} \qquad [46]$$

Where R_0 is the characteristic Forster distance where 50% efficiency is observed. Such a process requires that there is significant overlap between the spectra of donor and acceptor chromophores. Thus, if appropriately calibrated, the phenomenon of FRET can be used as a 'Spectroscopic nano-ruler'. In the specific context of protein folding, the native state is marked by a significant aggregation of the amino acid residues into a compact structure. Hence the FRET efficiency is quite high when the protein is in its folded state and low in the unfolded state, and there is an apparent instantaneous increase in the FRET efficiency when there is a transition from folded to unfolded state. This procedure has been used in most of the single molecule experiments, including the one we are concerned with currently. A brief description of the method used by Chung and Eaton is provided.

3.2 METHOD.

Chung and Eaton(1) used a confocal microscope system for their single molecule experiments, and used pulsed lasers to excite the donor chromophores. They obtained the donor and acceptor fluorescence signatures from the same objective. Proteins were immobilized on a glass slip with suitable binding agents and the chromophores are attached to the cysteine residues. Single photon avalanche diodes were used as detector to record individual photons' intensity, frequency and arrival time. A large number of 'photon trajectories' (explained in the next section) were collected using an automated data collection scheme(46). A brief description of the protein used is given below:

The Protein α_3D.

The protein used in this study is a 73-amino acid residue, designed polypeptide with 3 - helical motifs. The cysteine residues have been chemically modified to attach them to maleimide derivatives of Alexa TM dyes. The protein is

Fig. 10. Schematic depiction of H/D exchange taken from REF

immobilized on a polyethylene glycol surface with a biotin-streptavidin-biotin linkage, connected to the polypeptide through an AviTag TM linker and another auxiliary linker. At neutral pH, this protein is known to exist in only two states – folded and unfolded, and a glassy phase isn't observed at this condition.

4. Data Analysis

The information concerned with the folding process is obtained from the 'photon trajectories'. A sequence of photons along with their colour, signifying whether they are emitted by donor or acceptor chromophore, and the arrival time constitutes a photon trajectory. The protein molecule could be several 'states' or 'conformations' each of which is characterized by the number of acceptor and donor photons they emit.

One of the procedure to obtain the folding time from the photon trajectories is 'Time binning': The observation time is divided into small intervals or 'bins'. The number of donor and acceptor photon in that particular time bin is recorded and the fraction of acceptor photons is a measure of FRET efficiency and shall be referred to here on as 'apparent FRET efficiency':

$$E_i = \frac{n_{A_i}}{n_{A_i} + n_{B_i}} \qquad [47]$$

This FRET efficiency can then be attributed to the protein being in the native folded or the unfolded state. If the folding time is much larger than the bin time, then this method is quite reliable, but if it is lesser than the bin time, the estimation of folding time measurements becomes erroneous. Hence, there arises a necessity for an elaborate statistical method, more rigorous than this time binning technique. The maximum likelihood method for estimation of folding time, developed by Gopich and Szabo(47) is described here.

4.1 Maximum likelihood estimation.

This method is used to estimate the parameters that decide the probability distribution of a large data set, with the knowledge of the data values that a small subset of data points chosen at random takes, by maximizing the so called 'Likelihood function'. Stated roughly, we try to find that parameter or set of parameters, which would make the probability that the chosen data points acquires the specific values observed becomes maximum.

Assuming that a two state model describes the folding aptly, the parameters of the likelihood function are the 2 rate coefficients of transitions(folding to unfolding, unfolding to folding), and the apparent FRET efficiencies of the 2 states possible. These quantities are represented as matrices:

$$E = \begin{pmatrix} E_F & 0 \\ 0 & E_U \end{pmatrix}, K = \begin{pmatrix} -k_U & k_F \\ k_U & -k_F \end{pmatrix}, p = \begin{pmatrix} p_F \\ 1 - p_F \end{pmatrix} \qquad [48]$$

Where E_F and E_{UF} are FRET efficiencies in folded and unfolded states, P_{eq} is the equilibrium probability density in folded and unfolded states and k_F and k_U are rate coefficients of the transitions:

The appropriate likelihood function is set up recursively by tracking the modification of probability density of each of the states, with the emission of every photon:

$$L_j = 1^T \prod_{k=2}^{N_j} [F(c_k) exp((K\tau_k))] F((c_1)) p_(eq) \qquad [49]$$

The likelihood function is the probability of observing a particular photon trajectory. In the above equation, $\mathbf{K}$ is the rate constant matrix given by eq.(49), $\mathbf{F}(c_k)$ is the matrix which denotes the colour of the kth photon, thus $\mathbf{F}$(acceptor) = E and $\mathbf{F}$(donor) = $\mathbf{I}$ - $\mathbf{E}$. τ_k is the time between the $(k-1)^{th}$ photon and k^{th} photon. This function can be interpreted as follows: Before any photon is incident, the protein has an equilibrium probability density. This density is modified and is marked by the incidence of first photon. The leftmost terms denote this change. In the time between the arrival of the first and second photon, the protein evolves in time and the exponential term represents this evolution. This procedure is continued for the whole photon trajectory. It is then summed over all the states (in this case, it is a two state problem) which is represented by 1^T.

Another such likelihood function can be set up, when the number of photons emitted by each configuration/state isn't the same. For continuous transitions, this method can be used by discretizing the space accordingly. In practical uses, the logarithm(ln L) of the likelihood functions are maximized rather the likelihood function itself. This two state model assumes an instantaneous transition from the unfolded state to the native folded state, which is quite unlikely as it implies zero transition path time.

In a modified case, assuming a 3 state transition from unfolded state to a transient state to the native folded state, we can account for a finite transition time. The parameters are appropriately modified and the transition path time is obtained as the 'lifetime' of the intermediate state. The rate constant matrix in this case is:

$$K = \begin{pmatrix} -k_{U'} & k_S & 0 \\ k_{U'} & -2k_S & k_{F'} \\ 0 & k_S & -k_{F'} \end{pmatrix} \qquad [50]$$

The lifetime of the transition state, which can be interpreted as the average transition path time is hence:

$$\tau_S = 1/(2k_S) \qquad [51]$$

The accuracy of this method can be checked by plugging in the obtained parameter values into simulations and comparing the results obtained.

In the next section, we discuss the findings of Chung et al, where they hint at an apparent breakdown of Kramers' theory, and then claim that there could be the impact of 'internal friction' within the protein which causes the sub-linear relation between the inverse of diffusion coefficient with the coefficient of viscosity.

5. DISCUSSION

5.1 Deviations from Kramers' rate.

Kramers' theory hinges on a set of careful and generic assumptions such as the Brownian motion assumption, uniform diffusion coefficient assumption etc. In many physical cases,

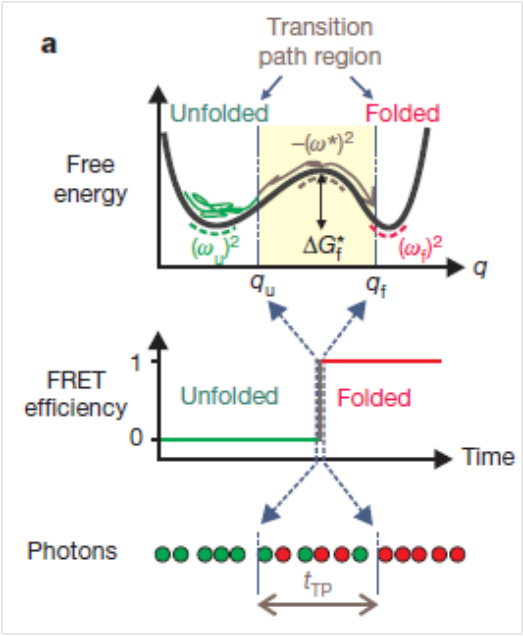

Fig. 11. Plots of difference between the log likelihood function for the 3 state model and 2 state model(delta ln L). As can be observed, there is a sharp spike at a characteristic transition path time, signifying the failure of the 2 state instantaneous transition assumption. Figure taken from Ref. (1)

memory effects of the solvent could be significant enough to make the brownian motion assumption of delta correlation fallible. For instance, such memory effects could result in considerable deviations from the usual Kramers' theory, but would be consistent if a frequency dependent friction is assumed(48). A generalised theory where the Langevin equation in Kramers' paper is replaced with the generalised Langevin equation, that has a time dependent friction, has been proposed(49) and had found considerable experimental evidence(48).

In such a scenario, where a multitude of factors could possibly contribute to deviations in Kramers' rate, it is necessary to find the attribute deviations to the right cause. Chung et al claim that there is but a sublinear dependence(1/D* proportional to $\eta^{0.3}$) of the inverse of diffusion coefficient on viscosity. This weak dependence possibly mean that the breakdown of the usual Kramers is perhaps not due to the frequency effects, but due to some internal effects within the protein, in such a way that the diffusing coordinate - the fraction of native residue(Q) couples with the solvent in a non-conventional way.

Friction, in the sense associated with diffusion in a solvent, arises due to the fact that the free particle motion in the solvent is arrhythmically hindered by random collisions with the solvent particles. However, in this case of protein folding, various non-native interaction forces, which are essentially random could also hinder the motion of the diffusing coordinate, apart from the solvent effects. This effect has been observed in studies concerned with protein conformational changes(50), where a part of the amino acid chain is exposed to the solvent atoms and the other part experiences only internal interaction forces. This internal friction, referred so since it has its origin from internal forces, can be, in principle accounted and modelled by a single ad hoc internal friction coefficient. Internal friction can be both 'wet' and 'dry'(51) : It can be due to solvent mediated effects or purely due to internal interaction forces. In such a case, we need a theory to describe ways to incorporate solvent effects into the internal friction coefficient. The internal friction arises from the fluctuating interaction forces, which had been earlier quoted to be responsible to the energy landscape roughness. Wolynes and Bryngelson(30) had derived an expression for the diffusion coefficient which exhibited super-arrhenius relation with temperature(eq.29), which was similar to the one derived by Zwanzig from a simple one dimensional model(29). This diffusion coefficient can then be related to the internal friction. Chung et al have claimed that such super arrhenius behaviour fits the data for temperature variation of transition path time very well too. But, a microscopic theory of internal friction is required for understanding the physical basis of internal friction. In the next section, we describe the extension of Rouse model(52) for polymer dynamics, to understand and quantify internal friction.

5.2 Rouse Model with Internal friction.

The internal friction coefficient can be derived from a microscopic picture by considering the polymer to be network of beads interconnected by springs, in the similar spirit as Rouse's Model(52). A langevin equation is set up, considering an internal frictional force which retards the relative motion between a pair of connected beads.

$$-\xi_S \frac{d\mathbf{r}}{dt} - \xi_i \mathbf{k} \frac{d\mathbf{r}}{dt} - k_0 \mathbf{k}\mathbf{r} + \mathbf{f}(t) = 0 \qquad [52]$$

Here, a harmonic potential is assumed for the spring connecting the beads. E_s denotes the solvent friction, E_i represents the microscopic frictional coefficient related to the retarding force which resists the relative motion between the beads, **K** is the spring constant Hessian matrix, **r** is the position vector of the beads and **f**(t) is the random force. In terms of the individual bead coordinates, the langevin equation is:

$$\zeta_s \frac{d\mathbf{R}_n}{dt} = (\kappa + \zeta_i \frac{\mathrm{d}}{\mathrm{dt}})(\mathbf{R}_{n-1} - 2\mathbf{R}_n + \mathbf{R}_{n+1}) \qquad [53]$$

From eq. 5, by obtaining the eigen values of the matrix -ξ_i***K** - k_0***K**. The eigen values of the K matrix alone would have yielded the collective frequency of the normal modes of the bead. The internal friction adds a small correction term to those eigen frequencies and the random force introduces a homogeneous fluctuation over the whole chain, which is physically relatable to the case of protein folding. The characteristic time is the inverse of the frequency of the eigen modes and is given as:

$$\tau^{(n)} = (\tau_{Rouse}/n^2) + \tau_i \quad (disc3) \qquad [54]$$

$$\tau_{Rouse} = \frac{\xi_S}{3\pi^2} \frac{N\langle|\mathbf{r}_N - \mathbf{r}_1|^2\rangle}{k_B T} \quad (disc4) \qquad [55]$$

Where N is the total number of beads, τ_n is the frequency associated with the n^{th} normal mode, and τ_{rouse} is the 'Rouse time' i.e the one corresponding to the fundamental mode frequency. Thus we have an simple additive relation for the relaxation time. The diffusion coefficient can be identified here as D= k_BT/(Nξ_S). It can be shown that the solvent relaxation time is linearly related to this relaxation time and the internal friction is independent of the solvent relaxation time. This has also been experimentally verified in myoglobin

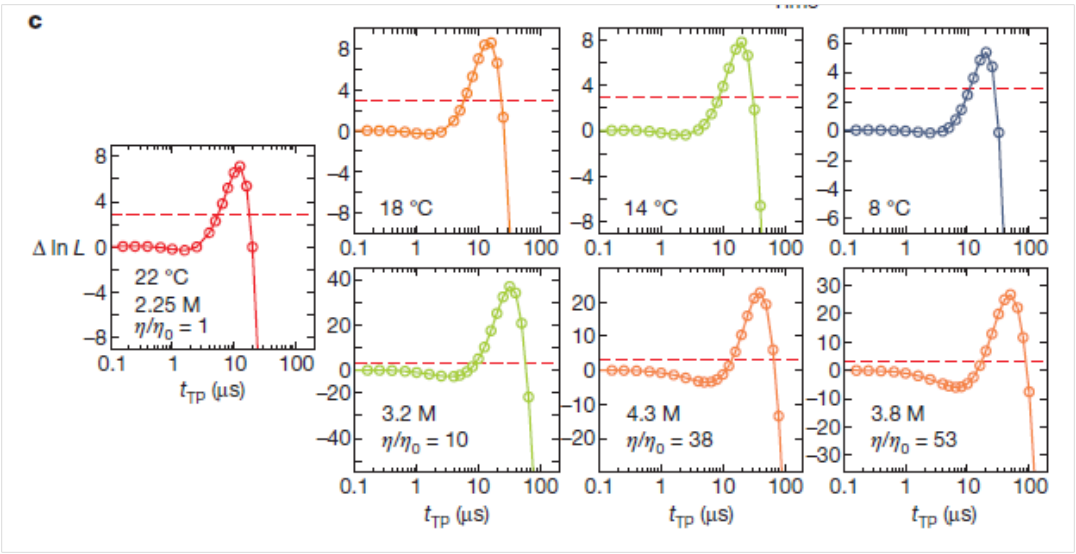

Fig. 12. A schematic representation of a free energy diagram, transition path region, FRET efficiency evolution and photon trajectory. Image taken from Ref.(1)

relaxation measurements. The theoretical model can be made further accurate for the specific case of proteins by considering the effects of chain compaction when denaturants concentration is reduced. But there are fundamental limitations to the theoretical estimates: Our assumption of solvent independent internal friction breaks down at high concentration of viscogens, when the overall topology of the free energy surface is altered. We shall describe a few key aspects of experimental quantification of internal friction in the next section.

5.3 Quantifying internal friction using experiments.

One of the simplest way to deduce the internal friction coefficient experimentally is by probing the internal friction by solving viscosity variation. At high denaturant concentration, the native interactions which are essentially responsible for the internal friction, are quite negligible. Thus, by estimating the chain reconfiguration times at various concentration of the denaturant in various solvent and then extrapolating it to the case when $\eta = 0$ yields the relaxation time at zero viscosity which can be solely attributed to internal friction. This extrapolation could well be erroneous and hence a more rigorous method is required to estimate the contributions from internal friction.

Alternatively, we can vary the segment length of the protein and then analyse the reconfiguration or relaxation time corresponding to each segment length. Since the relaxation time varies with the segment length as can be inferred from eq. 48, this could be an elegant to map out the relaxation time at various length and then extract the contribution from internal friction. This is carried out by attaching the donor and acceptor FRET labels at various positions of the bead and thus analysing the reconfiguration time. The information on chain dynamics is then obtained from nanosecond fluorescence correlation spectroscopy(nsFCS) (50) which allows dynamic monitoring of the fluctuations in the donor-acceptor distance.

To describe analytical expressions for these reconfiguration times, it is assumed that the donor and acceptor labels are attached as beads named 'D' and 'A' respectively, and that these beads are in turn connected to bead numbered i and j

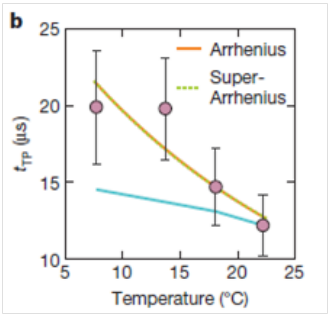

Fig. 13. The average transition path time is plotted as a function of the temperature. Both Arrhenius and Super-Arrhenius dependence of transition path time seem to fit the plot fairly. Figure obtained from Ref. (1)

respectively. Because of this, there is a small perturbation in the local potential energy

$$V_l = (1/2)k_l(\mathbf{r}_D - \mathbf{r}_i)^2 + (1/2)k_l(\mathbf{r}_A - \mathbf{r}_j)^2 \quad (disc5) \quad [56]$$

A 'friction scaled' Langevin equation is set up, wherein the internal friction effects are scaled to give a generalized equation

$$-\frac{d\widetilde{\mathbf{r}}}{dt} - \widetilde{\mathbf{K}}\widetilde{\mathbf{r}} + \widetilde{\mathbf{f}}(t) = 0 \quad [57]$$

Where $\widetilde{K}$ is the scaled Hessian matrix and other terms are scaled representation of the corresponding quantities. The relative distance between D and A can be written in terms of the normal modes of the scaled Hessian matrix

$$\mathbf{r}_{DA} = \mathbf{r}_D - \mathbf{r}_A = \frac{\widetilde{\mathbf{r}}_D}{\sqrt{\xi}_D} - \frac{\widetilde{\mathbf{r}}_A}{\sqrt{\xi}_A} = \frac{\sum_n x^{(n)}\widetilde{u}_D^{(n)}}{\sqrt{\xi}_D} - \frac{\sum_n x^{(n)}\widetilde{u}_A^{(n)}}{\sqrt{\xi}_A} \quad [58]$$

Where u_A and u_D are normal mode vectors corresponding to the acceptor and donor coordinates respectively. The reconfiguration time for this segment length r_D-r_A is then

$$\tau_{DA} = \frac{\int_0^\infty dt \, \langle \mathbf{r}_{DA}(0)\mathbf{r}_{DA}(t)\rangle}{\langle \mathbf{r}_{DA}^2(0)\rangle} \quad [59]$$

With the assumption of eq. (47) and equipartition theorem, the final expression for the reconfiguration time is

$$\tau_{DA} = \frac{\sum_n (\frac{\widetilde{u}_A^{(n)}}{\sqrt{\xi_A}} - \frac{\widetilde{u}_D^{(n)}}{\sqrt{\xi_D}})^2 (\tau^{(n)})^2}{\sum_n (\frac{\widetilde{u}_A^{(n)}}{\sqrt{\xi_A}} - \frac{\widetilde{u}_D^{(n)}}{\sqrt{\xi_D}})^2 \tau^{(n)}} + \tau_i \quad [60]$$

The unknown parameters in this context are the spring constants, friction coefficients of each of the beads and the internal friction relaxation time. The spring constants can be obtained from FRET efficiency histograms obtained from nsFCS, and the friction coefficients can be found out too.This leaves the only parameter τ_i to be estimated, which can then easily be extracted.

This method clearly has its own flaws too since we assume an oversimplified model for proteins which are bound to fail in near-native situations at low denaturant concentration. Nonetheless, this segment length analysis provides a fairly accurate paradigm to estimate the internal friction coefficient. In the final section, we end the discussion with a closing note.

Conclusion

A complete description of the protein folding problem with insights from biological, theoretical and experimental studies, along with methods of data analysis has been provided. Single molecule studies were used to study the folding of an alpha3D protein and a sublinear dependence of inverse of diffusion coefficient on viscosity was reported by Chung and Eaton. This was hypothesized to be due to the internal friction within the protein chain. However, there have been instances where an interplay of both non-markovian effects and internal friction were attributed to the reduced friction dependence(53). A generalized langevin equation was assumed to be describing the dynamics of the bead in Rouse model and a Grote-Hynes sort of a procedure is used to obtain expressions for the relaxation time due to internal friction. It has also been observed that increased frustration arising from non-native interactions can increase the effects of internal friction(54). Such observations call for renewed theoretical understanding and experimental studies.

Many fundamental biological questions can be answered with a concerted endeavour in both theory and experiments. Protein folding is still an open problem and there a lot of unanswered questions lingering. As Albert Einstein quote.

"The formulation of a problem is often more essential than its solution, which may be merely a matter of mathematical or experimental skill. To raise new questions, new possibilities, to regard old problems from a new angle requires creative imagination and marks real advances in science."

We can quite optimistically hope that new questions are raised, new possibilities are explored in this discipline and creative imagination in theory and experimentation keeps this field vibrant as it had been!

1. Chung HS, Eaton WA (2013) Single-molecule fluorescence probes dynamics of barrier crossing. *Nature* 502(7473):685–688.
2. Berg J, Tymoczko J, Stryer L (2002) Lipids and Cell Membranes. *Biochemistry* pp. 319–344.
3. Smyth M, Martin J (2000) X Ray Crystallography. *Journal of Clinical Pathology: Molecular Pathology* 53(1):8–14.
4. Wüthrich K (2001) The way to NMR structures of proteins. *Nature Structural Biology* 8(11):923–925.
5. Voet, Donald VJ (2014) *Biochemistry*. No. 1, pp. 1–5.
6. Greenfield NJ (2006) Using circular dichroism spectra to estimate protein secondary structure. *Nature Protocols* 1(6):2876–2890.
7. Edman P (1950) Method for determination of the amino acid sequence in peptides.
8. Levinthal C (1968) Are there pathways for protein folding? *Journal de Chimie Physique et de Physico-Chimie Biologique* 65:44–45.
9. Rooman M, Dehouck Y, Kwasigroch JM, Biot C, Gilis D (2002) What is paradoxical about Levinthal paradox? *Journal of biomolecular structure & dynamics* 20(3):327–329.
10. Bycroft M, Matouschek a, Kellis JT, Serrano L, Fersht aR (1990) Detection and characterization of a folding intermediate in barnase by NMR. *Nature* 346(6283):488–490.
11. John Ellis R (1996) Discovery of molecular chaperones.
12. Fersht aR (1997) Nucleation mechanisms in protein folding. *Current opinion in structural biology* 7:3–9.
13. Anfinsen CB, Haber E, Sela M, White FH (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America* 47(9):1309–1314.
14. Hunter P (2006) Into the fold. *EMBO reports* 7(3):249–252.
15. Muñoz V, Serrano L (1996) Local versus nonlocal interactions in protein folding and stability–an experimentalist's point of view. *Folding & design* 1(4):R71–R77.
16. Dill KA (1990) Dominant forces in protein folding. *Biochemistry* 29(31):7133–7155.
17. Dobson CM, Karplus M (1999) The fundamentals of protein folding: Bringing together theory and experiment.
18. Feynman RP (1972) *Statistical Mechanics*.
19. Bryngelson JDD, Wolynes PGG (1987) Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences* 84(November):7524–7528.
20. Derrida B (1980) The random energy model. *Physics Reports* 67(1):29–35.
21. Pande VS, Grosberg AYu, Tanaka T (1997) On the theory of folding kinetics for short proteins. *Folding & design* 2(2):109–14.
22. Wannier G (1950) The Triangular Ising Net. 79(2).
23. All UTC (2017) Annals of Mathematics Characteristic Vectors of Bordered Matrices With Infinite Dimensions Author (s): Eugene P . Wigner Source : Annals of Mathematics , Second Series , Vol . 62 , No . 3 (Nov ., 1955), pp . 548-564 Published by : Annals of Mathematic. 62(3):548–564.
24. Onuchic JN, Wolynes PG, Luthey-Schulten Z, Socci ND (1995) Toward an outline of the topography of a realistic protein-folding funnel. *Proceedings of the National Academy of Sciences* 92(8):3626–3630.
25. Landau LD, Lifshitz EM (1980) Statistical Physics Part I.
26. Plotkin SS, Onuchic JN (2002) *Understanding protein folding with energy landscape theory Part I : Basic concepts*. Vol. 2, pp. 111–167.
27. Luthey-schulten Z, Wolynes PG (1997) THEORY OF PROTEIN FOLDING : The Energy Landscape Perspective. (1):545–600.
28. Chan HS, KA (1998) Protein folding in the landscape perspective: Chevron plots and non-arrhenius kinetics. *Proteins: Structure, Function, and Genetics* 30(1):2–33.
29. Zwanzig R (1988) Diffusion in a rough potential. *Proceedings of the National Academy of Sciences of the United States of America* 85(7):2029–2030.
30. Bryngelson JD, Wolynes PG (1989) Intermediates and barrier crossing in a random energy model (with applications to protein folding). *The Journal of Physical Chemistry* 93(19):6902–6915.
31. Hänggi P, Talkner P, Borkovec M (1990) Reaction-rate theory: fifty years after kramers. *Reviews of modern physics* 62(2):251.
32. Socci ND, Onuchic JN, Wolynes PG (1996) Diffusive dynamics of the reaction coordinate for protein folding funnels. *The Journal of chemical physics* 104(15):5860–5868.
33. Klimov DK, Thirumalai D (1998) Lattice models for proteins reveal multiple folding nuclei for nucleation-collapse mechanism. *Journal of molecular biology* 282(2):471–492.
34. Wang J (2004) The complex kinetics of protein folding in wide temperature ranges. *Biophysical journal* 87(4):2164–2171.
35. Hummer G, Eaton W (2012) Transition Path Times for DNA and RNA Folding from Force Spectroscopy. *Physics* 5:1–3.
36. Hummer G (2004) From transition paths to transition states and rate coefficients. *Journal of Chemical Physics* 120(2):516–523.
37. Kim WK, Netz RR (2015) The mean shape of transition and first-passage paths. *Journal of Chemical Physics* 143(22).
38. Risken H (1989) The Fokker-Planck Equation: Methods of solution and applications. *Springer* p. 472.
39. Ben-Naim A (2013) Myths and verities in protein folding theories: From Frank and Evans iceberg-conjecture to explanation of the hydrophobic effect. *Journal of Chemical Physics* 139(16).
40. Bowman RM, Dantus M, Zewail AH (2013) Reprint of: Femtosecond transition-state spectroscopy of iodine: From strongly bound to repulsive surface dynamics. *Chemical Physics Letters* 589(43):42–45.
41. Zagrovic B, Pande VS (2004) How does averaging affect protein structure comparison on the ensemble level? *Biophysical journal* 87(4):2240–6.
42. Pocker Y, Biswas SB (1980) Conformational dynamics of insulin in solution. Circular dichroic studies. *Biochemistry* 19(22):5043–9.
43. Englander SW, Kallenbach NR (1983) Hydrogen exchange and structural dynamics of proteins and nucleic acids. *Quarterly Reviews of Biophysics* 16(04):521.
44. Moerner WE, Fromm DP (2003) Methods of single-molecule fluorescence spectroscopy and microscopy. *Review of Scientific Instruments* 74(8):3597–3619.
45. Piston DW, Kremers GJ (2007) Fluorescent protein FRET: the good, the bad and the ugly. *Trends in Biochemical Sciences* 32(9):407–414.
46. Chung HS, McHale K, Louis JM, Eaton WA (2012) Single-Molecule Fluorescence Experiments Determine Protein Folding Transition Path Times. *Science* 335(6071):981–984.

47. Gopich IV, Szabo A (2009) Decoding the pattern of photon colors in single-molecule FRET. *Journal of Physical Chemistry B* 113(31):10965–10973.
48. Velsko SP (1983) Breakdown of Kramers theory description of photochemical isomerization and the possible involvement of frequency dependent friction. *The Journal of Chemical Physics* 78(1):249.
49. Grote RF, Hynes JT (1980) The stable states picture of chemical reactions. II. Rate constants for condensed and gas phase reaction models. *The Journal of Chemical Physics* 73(6):2715.
50. Soranno A, et al. (2012) Quantifying internal friction in unfolded and intrinsically disordered proteins with single-molecule spectroscopy. *Proceedings of the National Academy of Sciences* 109(44):17800–17806.
51. Echeverria I, Makarov DE, Papoian GA (2014) Concerted dihedral rotations give rise to internal friction in unfolded proteins. *Journal of the American Chemical Society* 136(24):8708–8713.
52. Rouse PE (1953) A Theory of the Linear Viscoelastic Properties of Dilute Solutions of Coiling Polymers. *J. Chem. Phys.* 21(1953):1272–1280.
53. Makarov DE (2013) Interplay of non-Markov and internal friction effects in the barrier crossing kinetics of biopolymers: Insights from an analytically solvable model. *Journal of Chemical Physics* 138(1).
54. Sutto L, Lätzer J, Hegler Ja, Ferreiro DU, Wolynes PG (2007) Consequences of localized frustration for the folding mechanism of the IM7 protein. *Proceedings of the National Academy of Sciences of the United States of America* 104(50):19825–19830.