PHASE TRANSITIONS IN PROTEIN FOLDING: The Energy Landscape Perspective

Edgar Larios

May 2, 2000

Abstract

A quantitative understanding of protein folding is now possible with the help of the energy landscape theory and the protein folding funnel concept. Monte Carlo simulations show the correctness of this approach. In this paper it is presented an heuristic motivation of the energy landscape theory and the results obtained are probed using computational simulations.

1 Introduction

Proteins are chains of aminoacids. It is expected that we can distinguish proteins in function of the sequence of those amino-acids. Nevertheless X-ray diffraction revealed that proteins are not simple chains of structures as DNA. Actually, proteins are three-dimensional structures with complex folds. The 3-dimensional structure of the proteins are hard to predict a priori. It has been a puzzle how to predict the shape of a one-dimensional sequence of amino-acids when it becomes into a particular three-dimension object. This puzzle, the protein folding problem, has been studied by physical chemists, mathematicians and physicists.

Molecular scientists have seen protein folding as a complex chemical reaction. Another approach is to use statistical mechanics to understand the kinematics of protein folding. Therefore, it is possible to make an analogy with statistical physics in that folding resembles a phase transition in a finite system.In general, it has been hard to relate the folding with protein structure and energetics. One early attempt to approach this problem was to imagine that there exists a pathway for folding [1]. An alternative point of view is to assume that the the folding puzzle can be understood by viewing the overall surface energy of the protein. This approach is useful if we have an ensemble of structures rather than a few defined intermediate structures. If the first case is true, folding is a statistical physics problem. Then, an energy landscape approach can be used.The energy landscape theory comes from the notion that statistical characterization of the immense number of protein configurations is sufficient for understanding protein folding kinematics in many regimes. This statistical description has been developed using tools from the statistical mechanics of disordered systems, polymers, and phase transitions of finite systems.

In this paper the energy landscape approach is presented. In the first section an heuristic motivation is developed by Wolynes et al.[2]. In the second section, a diffusive model of folding is motivated and developed [3]. Numerical results are also presented which exhibits the correctness of this diffusion approach. In the third section a numerical simulation on the topography of the energy landscape is presented [4].

2 The protein folding energy landscape

In the simplest model one can assume that proteins are chains of monomers where there are not contact between them. In this simple picture the monomer's contribution to the protein's energy are random, and can be treated like those for a heterpolymer. In a more realistic model, where there are native contacts, the proteins are seen as minimally frustrated heterpolymers[5]. The energy of any compact conformation of such RHP is a sum of random interactions that give rise to a rough energy landscape [6]. Since the energy contributions can either be stabilizing or destabilizing, the RHP is a frustrated system. Thus the energy landscape is rugged. Biomolecular chains can sample many conformation during their motions

and have the possibility of making inappropriate contacts between residues. This is an effect of frustration [10]. Proteins obey a principle of frustration leading to a funnel-like aspect of the landscape.

Within each stratum we can define an average energy $\bar{E}(n)$. The late stages of folding will have few states similar to the native state. Because of the random interactions the density of states is approximately a Gaussian distribution with a variance ΔE. The thermally weighted probability is also a Gaussian distribution centered at the mean $\Delta E^2/2K_BT$. The entropy S is defined as $S(E) = K_B log[\Omega_0 P(E)]$ where Ω_0 is the number of the states of the polymer.

A complete and rigorous treatment of folding needs to know all the thermodynamic variables as a function of the order parameters. A useful order parameter is the fraction of native-like contacts Q. In particular, it is expected to know the average energy $\bar{E}(Q)$, the ruggedness $\sqrt{\Delta E^2(Q)}$, the density of states $\Omega(Q,E)$ or the entropy $S(Q,E)$, and the local glass transition $T_G(Q)$. Since the energy of a unfolded state comes from randomly contributions of many terms, the energy distribution of any stratum of the funnel is a Gaussian centered about the mean energy

$$P(Q,E) = \frac{1}{\sqrt{2\pi\Delta E^2(Q)}} \exp\left[-\frac{\left(E-\bar{E}(Q)\right)^2}{2\Delta E^2(Q)}\right] \tag{1}$$

In this REM model, the probabilities of any two states to have energy E_1 E_2 are uncorrelated, then joint probability is just the product $P(E_1)P(E_2)$. Also, if there exists γ configuration per residue, and we have a total of N residues, then the total number of configurations is $\Omega_0 = \gamma^N$. From Fig.1 it can be seen that as the protein is closer to the native state, the number of configurations decreases.

Let Ω_0 be the number of configurations with similarity measure Q with respect to the native structure and $S_0(Q)$ the entropy. The density with energy E and similarity Q is then

$$\Omega(Q,E) = \Omega_0(Q)P(Q,E) \tag{2}$$

and the total entropy is

$$S(E,Q) = S_0(Q) - k_B\frac{\left(E-\bar{E}(Q)\right)^2}{2\Delta E^2(Q)} \tag{3}$$

The free energy of the misfolded structures at similarity Q and temperature Tcan be readily written

$$F(Q,T) = E_{m.p.}(Q) - TS(E_{m.p.},Q) \tag{4}$$

In equation 7, $E_{m.p.}$ stands for the energy state most probable. The energy of this state can be calculated from the Boltzmann distribution in Fig.2

$$E_{m.p.} = \bar{E}(Q) - \frac{\Delta E^2(Q)}{k_BT} \tag{5}$$

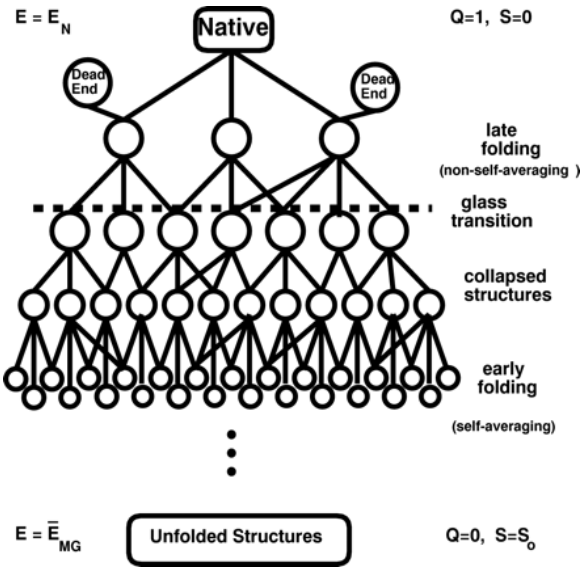

Figure 1: A thermodynamical pathway of folding toward the native state. Note that the number on available conformations reduces as Q increases. After a glass transformation protein can reach few states very of them are dead end states. Taken from [2] page 555.

The number of occupied states and entropy at this most probable energy are

$$\Omega(E_{m.p.}, Q) = \exp\left[\frac{S_0}{k_B} - \frac{\Delta E^2(Q)}{2K_BT}^2\right] \tag{6}$$

$$S(E_{m.p.}, Q) = S_0(Q) - \frac{\Delta E^2(Q)}{2k_BT^2} \tag{7}$$

So, equation 7 can be written as follows

$$F(Q,T) = \bar{E}(Q) - \frac{\Delta E^2(Q)}{2k_BT} - TS_0(Q) \tag{8}$$

When the system is at certain thermodynamic conditions, folding can be seen as a two steps reactions. Under this circumstance, the free energy shows two minima. One is near the folded state $Q \approx 0$, the other is near the unfolded state $Q \approx 1$. The later corresponds to a ensemble of folded states. The presence of the barrier involves the polymer physics on entropy loss and perhaps explicitly cooperative many body-forces for real proteins. Since the folded states have a much smaller configuration number, we can neglect the entropy of this folded state. Thus, as a first approximation, the free energy is equal to its interval energy E_n. Let's assume also that at the folding temperature, T_f, the probability of being in a folded state is equal to the probability of being in an unfolded state. That is

$$F_{native} = F(Q_{min}, T_F) \tag{9}$$

So, we can immediately obtain an expression for the slope of the funnel

$$\frac{\delta E_s}{T_F} = S_0 + \frac{\Delta E^2(Q_{min})}{2k_B T_F^2} \tag{10}$$

Where $\delta E_S = \bar{E}(Q_{min}) - E_N$ is the so called stability gap. A phase transition occurs at temperatures where there are too few states available, then, the system stays frozen in any of the few states. This fact is characterized by an entropy equal to zero for each stratum, i.e. $S(T_G, Q) = 0$. Using equation 10, we obtain the local glass transition temperature

$$T_G = \sqrt{\frac{\Delta E^2(Q)}{2k_B S_0(Q)}} \tag{11}$$

This results are summarized in fig. 2.

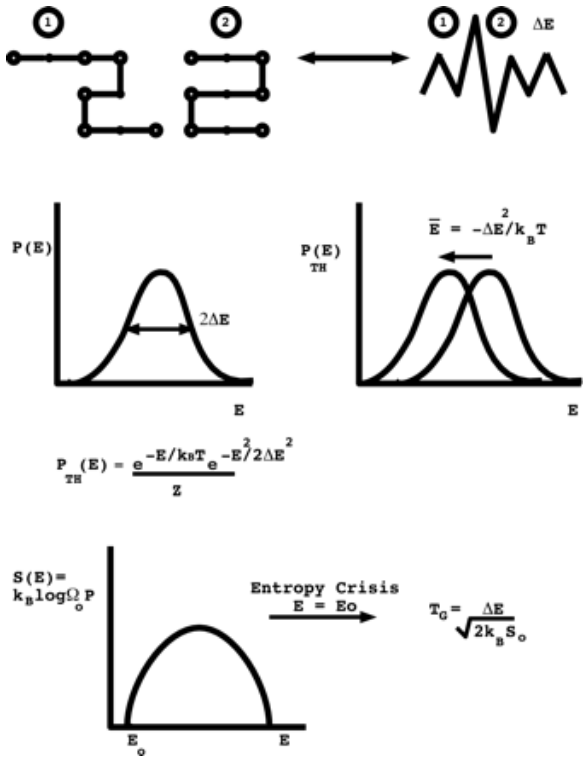

Figure 2: An sketch of random heterpolymer energy landscape. At the top we wee two uncorrelated lattices with different residues. At the center, it is shown that the density of states are a Gaussian distribution. At the bottom the energy of the system falls below E_0 and the entropy drops to zero. A glass transition is achieved at T_g. Image taken from [2] page 551.

A phase diagram is useful to show all the possible folding scenarios for a protein. In the Fig. 3 it is shown a phase diagram for different temperatures and Q as a

coordinate parameter. This is the energy landscape obtained from the minimally frustrated REM analysis. Note that the figures shown are just one slice over some average of the hydrophobicity of the sequence.

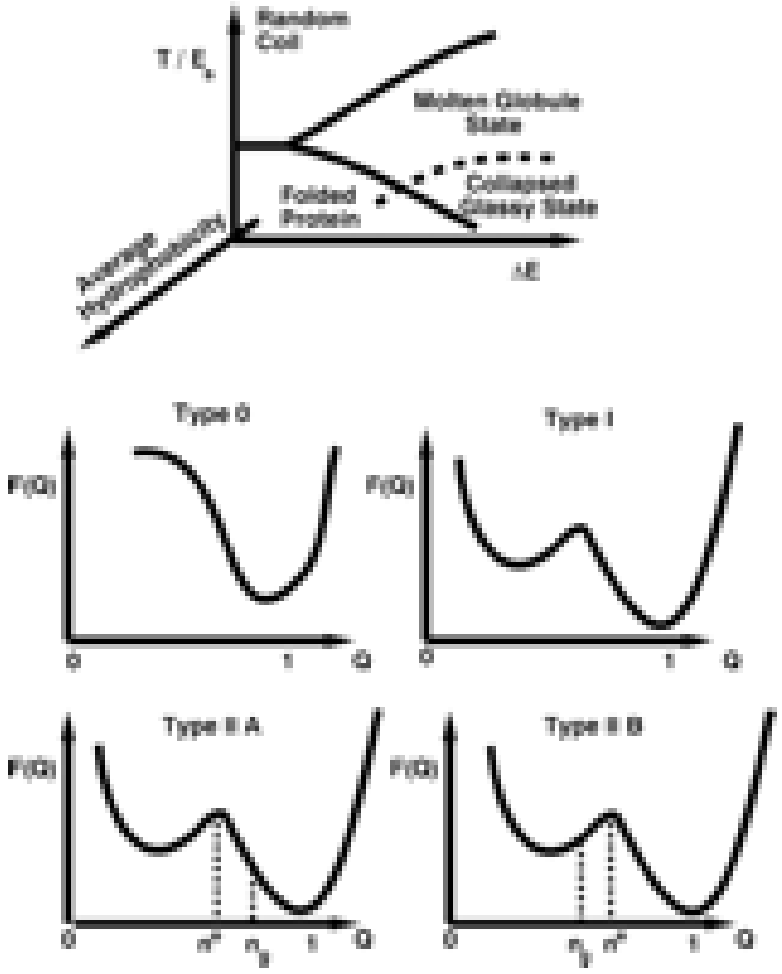

Figure 3: Phase diagram for folding as predicted by REM analysis. Image taken from [2] page 557.

Obviously it can be chosen another coordinates to represent the pase diagram. Simulations performed by Soccci & Onuchic on protein-like chains gives the same qualitatively picture [7]. This free energy diagram can be used to understand folding in different situations. It can be seen that in general the free energy can be either unimodal or bimodal. The first case is called Type 0 scenario. The later is called Type I scenario. In A Type 0A scenario there is not glass phase transition and the folding is a downhill process. Whereas in a Type 0B scenario there is a phase transition before complete folding; this phase transition determines the rate of folding. The scenarios Type I and Type 0 are usually found in the left hand of the phase diagram where the system is at temperature that correspond to the

native folded state $Q = 1$. In the right-hand part we find Type II scenario near to the coexistence region of folded, collapsed and collapsed frozen state. We can distinguish between scenario TypeII A and Type II B, in the later a phase transition occurs after the thermodynamic barrier, in a Type IIB the transition occurs before the protein reaches the barrier

3 Diffusive dynamics of the reaction coordinate

As it was mentioned above, folding occurs through a pathway formed by an ensemble of routes down a funnel energy landscape. This process is characterized by a formation of partially ordered structures. When the system flows through this energy landscape it is trapped by local minima. The different ensembles can be described by orders parameters or reactions coordinates. This funnel is characterized by a free energy that is function of the order parameters. The free energy is going to be formed by the energy and entropy of the system. the idea of this section is to study the features of this funnel by looking at the free energy as the the reaction coordinate changes. As the entropy of the funnel decreases so does the energy. The gradient of the free energy will distinguish between the average drift up and drift down of the funnel. Onuchic et al. study this process as an stochastic motion whose statistics depend on the jumps between local and minima. This process, as a good approximation, can be described as diffusion.

The energy landscape is stratified. That is, the statistical characteristics of the landscape depends on the distance from the native state. This similarity measure is called an order parameter in the physics of phase transitions. In the case of small systems such as proteins, the order parameter can be used as a reaction coordinate to compute the folding rate. In this section this reaction coordinate is going to be called n. (See Fig. 3).

Bryngelson and Wolynes in the late 1980's showed that to a first approximation, the folding rate can be calculated by grouping states with the same initial reaction coordinate in one stratum [8]. Then, it is possible to obtain a diffusion equation for the probability under the assumption that the reaction coordinate can change through small steps.

$$\frac{\partial P(n,t)}{\partial t} = \frac{\partial}{\partial n}\left\{ D(n)\left[\frac{\partial P(n,t)}{\partial n} + P(n,t)\frac{\partial \beta F(n)}{\partial n}\right]\right\} \tag{12}$$

The diffusion coefficient, $D(n)$, reflects the ruggedness of the energy landscape. Note that the average direction of the flow points in the same direction that the gradient of the free energy. In this model, the free energy is made of two terms, the energy that is decreased when the system is approaching the native state, ad the entropy term $-TS(n)$ which decreases with unfolding. Then, the free energy is strongly dependant on the temperature. At high temperatures, folding is suppressed since it is a thermodynamically uphill process. At the folding temperature the free energy

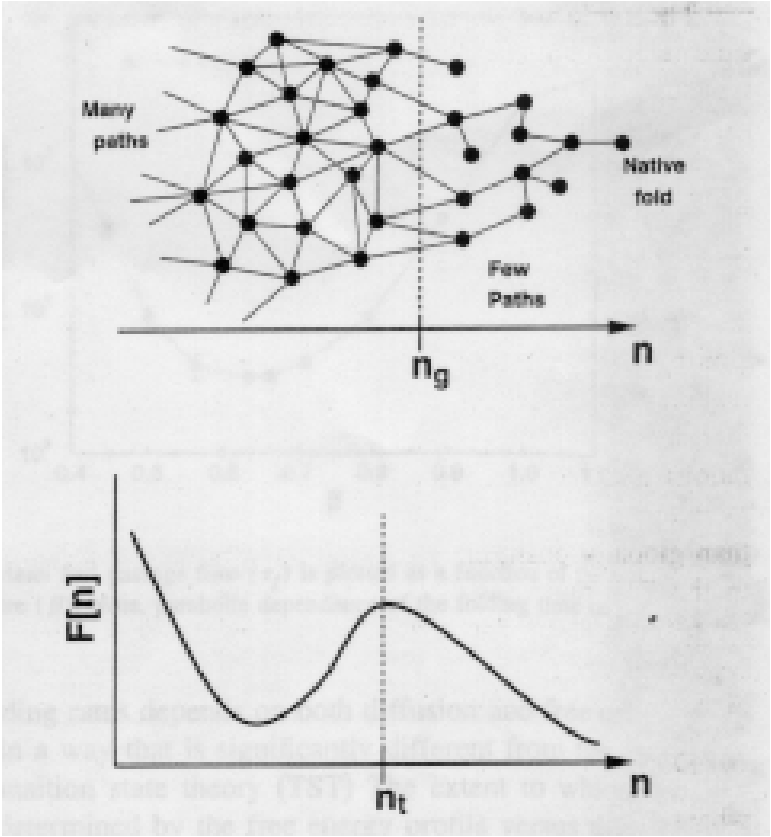

Figure 4: The top figure shows the behavior of the system as function of the arbitrary reaction coordinate n. n_g represent a transition point. The bottom image shows the free energy of the same system as function of the same arbitrary coordinate. Taken from [3] page 5861.

have two minima with a small thermodynamic barrier due to the no cancellation of the entropy by the energy as the systems moves through the funnel. At low temperatures folding is a downhill process. The folding time can be written as a double integral

$$\tau_f = \int_{n_{unf}}^{n_{fold}} dn \int_0^n dn' \frac{\exp\left(\beta F(n) - \beta F(n')\right)}{D(n)} \tag{13}$$

If there exists a barrier like in T_f, the free energy is bistable with one bottom close to n_{fold} and the other close to n_{unf}. If that is the case, the double integral can be approximated by a Kramers like law [9],

$$\tau_f = \left(\frac{2\pi}{\beta}\right)^{1/2} \frac{1}{D_0 \omega_{unf} \bar{\omega}_{fold}} \exp\left\{\beta[\bar{F}(n_t) - F(n_{unf})]\right\} \tag{14}$$

where $\bar{F}(n) = F(n) - TX(n)$ and $X(n) = log\left[\frac{D(n)}{D_0}\right]$. ω_{unf} is the curvature around n_{unf} and the $\bar{\omega}_{fold}$ is the curvature at the top of the barrier. D_0 is the effective

diffusion coefficient that the funnel will have in an flat landscape. The prefactor observed in equation 14 reflects the dependence of the time folding with temperature. This is evidence that the multiple recrossing occurs through diffusion. This situation is even more evident when the process is entirely downhill. In that case the double integral is proportional to the diffusion coefficient and barely depends on the slope of the free energy gradient. The diffusion coefficient depends on the topography of the energy surface of the protein and on the local displacements allowed to the protein. In the model presented here, the coefficient has a very strong dependence to the temperature that comes from the necessity to escape from local minima on the rough energy landscape.As it was shown by Wolynes et al.[3], at high temperatures D follows a Ferry law typical of glasses

$$D(T,n) = D_0 \exp[-\beta^2 \Delta E^2(n)] \tag{15}$$

At intermediate temperatures the diffusion coefficient shows a moderate non-Arrhenius behavior

$$D(T,n) = D_0 \exp[-S^*(n) + [\beta_g(n) - \beta]^2 \Delta E^2(n)] \tag{16}$$

This equation is valid for temperatures between T_g and $2T_g$. $S^*(n)$ is the configurational entropy at n. We can see that at at T_g the diffusion coefficient reduces abruptly by a factor of the total number of configurations. This fact may produce the so-called Levinthal paradox. Levinthal showed using a simple counting argument that a typical protein has far too many conformation to permit a thorough search for the global minimum. Below T_g we still expect a diffusion coefficient dependent of the activation energy. Nevertheless this activation energy will be very sensitive to details of the model and to the sequence of the heterpolymer. Then, in this regime of temperatures, ideas of statistical physics usually are not used and the slow folding process is better described by few kinetic pathways. Both the temperature dependence of D and the role of entropy and energy produces a parabolic dependence of the folding time with β. The agreement between theory and simulations is remarkable . Therefore, the model presented is not only qualitatively correct but also gives a fairly good prediction of the folding time for temperatures above T_g (see Fig.5). It also suggests that the folding kinetics can be represented as a diffusive process in the Q coordinate.

This result shows that the folding rates depend on both diffusion and free energy in a different way from the standard transition state theory (TST). The extent that the time of folding depends on the free energy and the diffusion coefficient is function of the choice of the reaction coordinate.

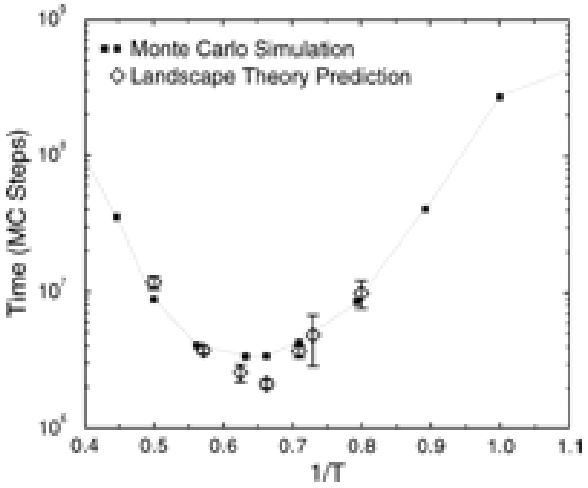

Figure 5: Comparison of the time folding obtained from Monte Carlo simulation with the theoretical predictions using the landscape theory. Taken from [2] page 579.

4 An outline of the topography of a realistic protein-folding funnel

The energy landscape philosophy and the analogy to phase transitions has provided the best approach to the protein folding problem. As it was said earlier on this paper, the energy landscape is based in a law of corresponding states mapping the phase diagram and kinematics mechanisms of real proteins onto simple models. At the empirical level, this law is named as the law of corresponding states and it is also the basic idea of the renormalization group. In a more complex phase diagram, where more degrees of freedoms are presented, the effect of extra degrees of freedom will be to "renormalize" energy and proteins scale for the funnel. Based on simulations made by Onichic et al. [4] it is possible to obtain the figure 6. This protein funnel scape is shown in scale. In the picture the width represent the entropy and the depth represent energy and two correlated structure scales. The barrier heights of the funnel is represented by ΔE and the energy of the folded state is scaled to the depth of the diagram.

The Molten Globule Band represents a transition region that acts a bottleneck. The coordinates that are examined are the fraction of angles in their native configuration, A, and the fraction of native contacts, Q. Note that the coordinate A changes by an small amount in comparison with Q. That suggests that A changes by a very small amount in each computational step, then gradients of free energy with respect to A are negligible. Since Q is highly related to the the interaction energy function, it is very useful to describe overall topography. In contrast wit A, gradients of free energy with respect to Q can be very large. The simulation made represent a time folding $\tau_0 \approx 3msec$. The sequence shown was made with

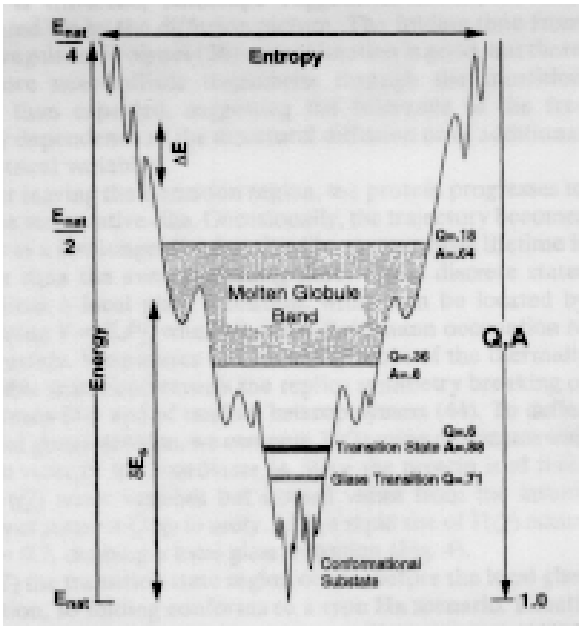

Figure 6: Schematic funnel for a realistic 60 amino acid helical protein. The coordinates Q and A have been normalized to their maxima values. Taken from [4] page 3628.

$\approx 3 \times 10^6$ steps. To obtain the density of states as a function of the energy and order parameters it was used a Monte Carlo histogram technique [10]. Using the density of states, Onuchiic et al. were able to determine the local thermodynamic glass transition, and the folding temperature.

In the Fig. 7, the free energy is plotted as function of Q and A at T_f. The free energy is bistable. The free energy has a saddle point at $A \approx 0.88$ and $Q \approx 0.6$. That is, each native contact is made during three fifths of the time in an ensemble of configurations of the transition state. This result is in agreement with the experimental value that shows that the transfer coefficient of mutations at each site ϕ varies between 0.3 and 0.7 [11]. Since the free energy in the region shown is flat, the thermodynamic barrier is small but broad. Therefore, there are multiple recrossings of the transition state region caused by the landscape ruggedness. Then as it was suggested lines above, folding times must be computed using a diffusive picture instead of a standard transition state which neglects recrossings. By looking at the correlated fluctuations of the coordinates Q and A it is obtained the diffusion constants for Q in the Molten Globule region $D \approx 3.5 \times 10^6$. Assuming that the free energy is harmonic and that the barrier top curvature is equal to the wall's equation 14 reduces to

$$\tau_f = 2\pi\tau_{corr} \exp F^*/k_B T \tag{17}$$

where F^* is the activation barrier of $2.4 k_B T_f$ from the two-dimensional plot and τ_{corr} is the correlation time for the harmonic fluctuations. In this simulation, τ_{corr} for both

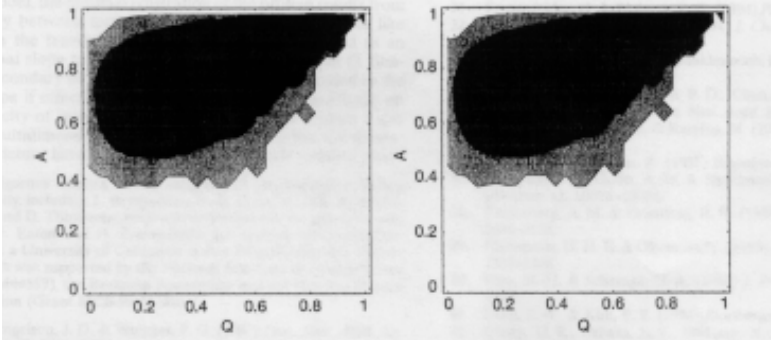

Figure 7: Two transition regions plotted onto the $Q - A$ plane. Taken from [4] page 3629.

Q and A is about 20,000 time steps. The value obtained, $\tau_f \approx 1.4 \times 10^6$ is a little bit shorter that the simulated value. The Bryngelson-Wolynes approximation is good, but the discrepancy obtained suggests the relevance of the additional geometrical variables or even the frequency dependence of the structural diffusion. The landscape ruggedness is well accounted by the diffusion picture since the system becomes glassy only after the transition region is crossed. The protein gets more native-like when the transition region is traversed. As it was mentioned earlier, the protein gets caught in dead-end states whose lifetimes is shorter than the average folding time. These discrete states arise from a local glass transition which can be located by computing

$$Y = \sum_i P_i^2 \tag{18}$$

where P_i is the Boltzmann occupation number of a microstate. In this picture, Y measures the inverse number of the thermally occupied states and shows the symmetry breaking of spin glasses [12]. Onuchic et al compute $Y(Q)$ for states with a given value of Q (see Fig. 8). At T_f there is an abrupt change in $Y(Q)$ at $Q \approx 0.7$ which defines a local phase transition. (See Fig. 8). Note that at T_f we have a Type IIA scenario since the transition state region occurs before the local glass transition.

5 Conclusion

The original approach to protein folding has changed in the last decade. An early attempt in which folding was view as a normal reaction in organic chemistry has evolve to a chemical physics viewpoint. This view has required a new set of theoretical ideas and computational techniques. The energy landscape theory provides a theoretical framework in which a full understanding of folding requires a global overview of the landscape. There is not an only pathway but an ensemble of routes toward the native state of the protein. Late in the folding process a protein may be trapped in single pathway but at this stage the protein has found

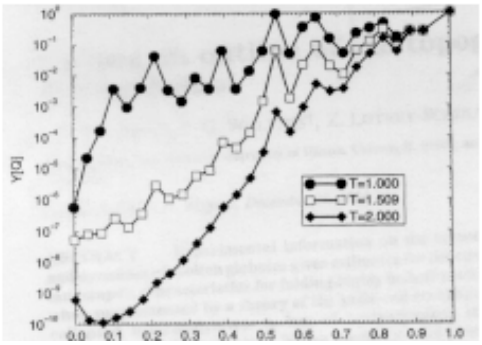

Figure 8: A plot of $Y(Q)$ as it is defined on the text. $Y(Q)$ is shown in three different temperatutes. Taken from [4] page 3630.

its correct folding configuration. Folding can be seen a process with a highly non-Arrhenius temperature dependence. This behavior is arise from the competition between trapping in misfolded states and drift down the folding funnel, this is a competition between energy and entropy. These effects can be described trough collective coordinates. The result presented here suggests that one coordinate is enough to describe the folding of one small lattice models over a wide range of temperatures.

References

[1] Kim PS, Baldwin RL.*Annu. Rev. Biochem.*, **59**:631, 1990.

[2] Onuchic JN, Luthey-Schulten Z, Wolynes PG.*Annu. Rev. Phys. Chem.*, **48**:545, 1997.

[3] Socci ND, Onuchic JN, Wolynes PG.*J. Chem. Phys.*, **104**:5860, 1996.

[4] Onuchic JN, Wolynes PG, Luthey-Schulten Z, Socci ND.*Proc. Natl. Acad. Sci. USA*, **92**:3626, 1995.

[5] Bryngelson J, Wolynes PG.*Proc. Natl. Acad. Sci. USA*, **84**:7524, 1987.

[6] Plotkin S, Wang J, Wolynes PG.*J. Chem. Phys.*, **106**:2932, 1997.

[7] Socci ND, Onuchic JN.*J. Chem. Phys.*, **103**:4732, 1995.

[8] Bryngelson J, Wolynes PG.*J. Phys. C*, **93**:6902, 1989.

[9] Onuchic JN, Wolynes PG.*J. Phys. Chem*, **92**:6495, 1988.

[10] Ferrenberg, AM, Swendsen RH.*Phys. Rev. Lett.*, **61**:2635, 1988. Ferrenberg, AM, Swendsen RH.*Phys. Rev. Lett.*, **61**:2635, 1988.

[11] Otzen DE, Itzhaki LS, Elmasry NF, Jackson SE, Fersht AR.*Proc. Natl. Acad. Sci.*, **91**:10422, 1994.

[12] Mezard M, Parisi G, Virasoro MA.*Spin Glass Thoery and Beyond*, **World Scientific, Singapore**, 1987.