UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL ATOMIC ENERGY AGENCY
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

H4.SMR/916 - 34

SEVENTH COLLEGE ON BIOPHYSICS:

Structure and Function of Biopolymers: Experimental and Theoretical Techniques.

4 - 29 March 1996

The Energy Landscape Theory of Protein Folding

Peter G. Wolynes
School of Chemical Sciences
University of Illinois
Urbana Illinois 61801
U.S.A.

The Energy Landscape Theory of Protein Folding

Peter G Wolynes and Zan Luthey-Schulten

School of Chemical Sciences, University of Illinois, Urbana, Illinois 61801

Abstract. Recent Progress in Protein Folding: Applying the energy landscape analysis to realistic protein simulations and structure prediction.

1 Introduction

Prediction of a protein's structure and folding mechanism from its sequence has been described as the determination of the second half of the genetic code (Gierasch and King (1990)). Proteins and nucleic acids are the simplest information bearing components of biological systems to have individual identities. Although the ongoing effort to map the entire DNA genome of several species including man has provided biologists with more than a hundred thousand protein sequences, the structures of only a few thousand are known. More importantly the rules or energy functions to turn these pieces of one dimensional information (sequences) into three-dimensional structures (folded proteins) are just becoming clear. Proteins are made up of amino acids that can have several conformations so that the folding process takes place on a rough multi-dimensional potential surface. Our understanding of the physics of protein folding is impeded by the complexity of the process. On the one hand, many of the features of protein folding dynamics are like those for any random heteropolymeric system. The folding route is not unique and passes through numerous misfolded structures whose structures and energies are unrelated. While on the other hand, protein sequences have evolved to allow proteins to fold to unique native states with certain local structural motifs and to carry out selective functions. For example the sequence of the oxygen carrying molecule, myoglobin, encodes the structure shown in Fig.1. It has a high degree of symmetry built up from the repetitive local helical units. How does this symmetrical structure come about? Why is the folding and unfolding of a protein apparently a reversible process?

The scientist seeking to answer these questions, must understand the properties of proteins as complex heterogeneous systems that can be thought of as disordered, but equally must try to uncover those nonrandom features of proteins that are essential for their folding, both in the sense of general themes and of detailed design. Although conventional potential functions have been developed to describe the interactions in a protein near its native state, they cannot yet be practically applied over the millisecond time scale needed to simulate the folding routes and kinetics. This seemingly hopeless situation is reminiscent of a similar challenge facing physicists in the days before BCS theory was invented to explain

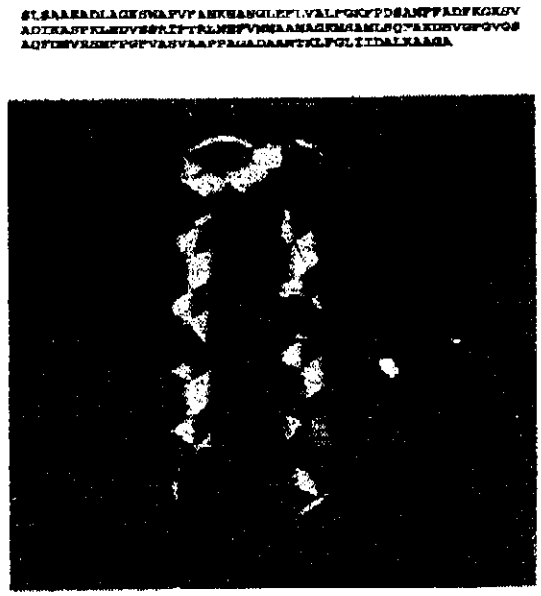

Fig. 1. Sequence and structure of myoglobin

superconductivity. Physicists were faced with the difficulty of describing a complex effect on very different dynamic scales: Superconductivity involves an effect that was on the level of 10^{-7} eV when the energies of electronic structures were only known to 0.1ev. This gave Bardeen, Cooper and Schrieffer license to craft a phenomenological theory of superconductivity that was then parametrized and verified by experimental data. Following their example we have attempted to describe the diverse behavior associated with protein folding using the statistical energy landscape approach, a phenomenological theory requiring only a few energy parameters in the simplest form: ΔE_s, the stability gap between the ground (native or folded) state of the protein and the mean of the excited (misfolded) states and δE, the roughness of the energy landscape. This approach has had considerable success in guiding us to consider good order parameters and appropriate collective reaction coordinates to interpret thermodynamic and kinetic protein folding experiments and suggesting detailed but simple energy functions for protein structure prediction (Friedrichs et al. (1991), Bryngelson et al. (1995))

The lectures here in Denmark will cover the basics of the statistical energy landscape approach, use it to interpret what are common and specific features observed in protein folding experiments, and describe how it leads to optimized energy functions for protein structure prediction. Reviews that deal with the energy landscape and other approaches to these topics can be found in (Bryngelson et al. (1995), Dill et al. (1995), Garel et al. (1995))

2 The Protein Folding Energy Landscape

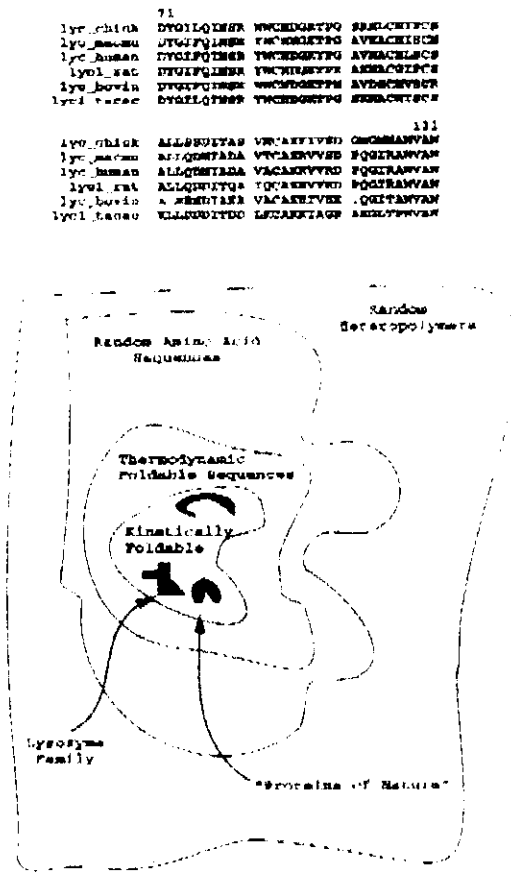

Fig. 2. a. Family of lysozyme sequences and their respective percent identities. b. How Proteins of Nature are embedded in the ensemble of random heteropolymers

Are proteins random objects? In Fig.2a are sequences of the protein lysozyme from various species. Even though it has essentially the same structure and function in all species, the sequence identities compared to the protein in the chicken are relatively low. Sections of the viral lysozyme are only 30% identical with the chicken, sequentially. Two English texts with only 30% identity would appear totally unrelated strictly orthographically, but could have the same meaning. Likewise proteins with structural homologs with such low sequence identities are not uncommon. The difficulty in extracting the meaning from protein sequences is in discerning what features are common to all sequences, what features are specific

to protein-like sequences, and which ones are specific to a given structure. How the ensemble of protein-like sequences is embedded in the ensemble of random heteropolymers is sketched in Fig.2b. Within the space of random heteropolymers based on the twenty naturally occuring amino acids, we need to further differentiate between thermodynamically foldable sequences and the subset of kinetically foldable sequences that make up the proteins of Nature. Families like the lysozyme sequences belong in this last category. Proteins of Nature, while only marginally stable and easy to denature with either heat or pH, must fold on a time scale that is relevant for the biological processes occurring in cells. This time is relatively short, less than a minute, which seems paradoxical given the large number of conformations that a protein can theoretically be in during folding. Quantitatively this can be understood by studying the formation of local structure, such as helix formation in collapsed polymers and understanding the funneled nature of the landscape for topological rearrangements (Luthey-Schulten et al. (1994))

2.1 Energy Landscape of a Random Heteropolymer

Many aspects of the folding process can be understood from studying the energetic properties of a random heteropolymer (RHP) using lattice models. The simplest of these makes use of two kinds of residues (e.g. hydrophobic and hydrophilic amino acids) randomly distributed. This is a useful model for visual illustration although it has some special properties that make it different from the more general 20 amino acid case. Both from theoretical calculations and simulations we know two basic facts about the random heteropolymer: a. Modest structural change gives rise to large change in energy. b. Low energy states are very different in structure.

In the general case the energy is a sum of random interactions that give rise to a rough energy landscape like the Alps. Since the energy contributions can either be stabilizing or destabilizing, the RHP is a frustrated system. In a leap of faith Bryngelson and Wolynes (Bryngelson and Wolynes (1987)) applied already in 1987 the random energy model (REM) developed by Derrida to describe spin glass systems to proteins (biopolymers), in particular to the misfolded states of a protein. The basic validity of this approach has since been borne out by numerous analytical and numerical studies making the REM to the zeroth order approximation for understanding biopolymers. A brief review of the basic features of the REM for the RHP is shown in Fig 3. As a result of the random interactions the density of states is approximately a gaussian distribution with a variance ΔE. The thermally weighted probability is again a gaussian distribution centered about the mean $E = -\Delta E^2/2k_BT^2$. The density of states can not really be gaussian in its tail, but runs out of low energy states. The entropy S is defined

$$S(E) = k_B log \Omega_c P \tag{1}$$

where Ω_c is the number of conformational states of the polymer. As the system is cooled the energy falls. The system runs out of entropy when the average

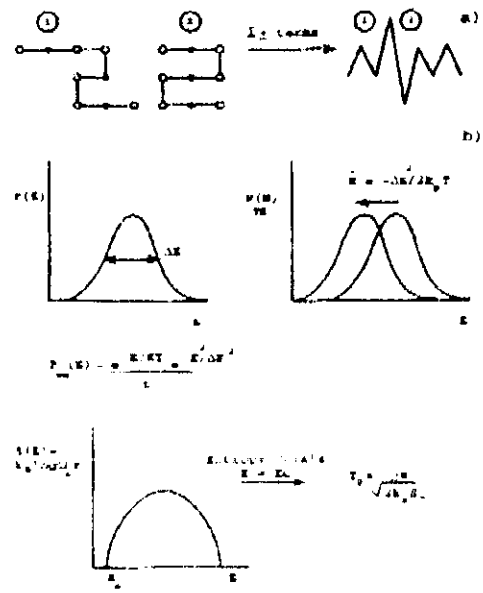

Fig. 3. a) Flat rugged landscape of RHP. b) Thermodynamics of REM approximation applied to a RHP

energy falls below a critical value $E \leq E_o$ such that $S(E_o) = 0$. This entropy crisis occurs at a temperature T_g where

$$T_g^{-1} = \sqrt{2k_B S_o/\Delta E^2} \tag{2}$$

where $S_o = k_B \log \Omega_o$. Below T_g the kinetics of the system exhibits glassy-like behavior depending on history. Above T_g the system behaves like a viscous liquid. Transition rates between different low energy states leads to a strong generally non-Arrhenius temperature dependence of the rate of exploring configuration space. A more detailed discussion of the kinetics on a REM landscape is given below.

The low entropy at glass transition makes it tempting to identify T_g with the folding temperature. Indeed Shakhnovich and Gutin have estimated that a large fraction of random sequences would have unique thermodynamically stable native states below T_g. The fraction of these sequences with significant thermal occupation of the native state is independent of chain length and is given by (Shakhnovich and Gutin (1990))

$$Prob(N.S.) = \frac{\sin(\pi T/T_g)}{\pi T/T_g} e^{-T/T_g}, T < T_g \tag{3}$$

where ϵ is related to the Boltzmann probability of the ground state E_o, $p_o = \exp(-E_o/k_B T)/Z > 1 - \epsilon$. Most of these sequences would still fold too slowly

in this temperature range because of glassy dynamics, i.e. trapping in other collapsed configurations. Instead we must examine the exponentially small fraction which can fold above T_g, and ask what is the simplest energy landscape that these sequences would have.

2.2 Simplest Viable Protein Folding Landscape

Postulate A: The energy landscapes of proteins are rugged because of the possibility of making inappropriate contacts between residues. It is reasonable to assume that when non-native contacts are made the energy contributions are random, and these contributions to the protein's energy can be treated just as for a random heteropolymer. In the ensemble of misfolded states with little native structure, the energetics can be described crudely by the REM shown in Fig 3. Low energy structures will appear unrelated and conformational changes are associated with a fluctuation $\sqrt{\Delta E^2}$ in the energy.

Postulate B: The Principle of Minimal Frustration There is a smooth overall slope to the energy landscape because of harmonious cooperativity. Native contacts and local conformational energies are more stabilizing than expected. This more realistic model considers the protein to be a "minimally frustrated heteropolymer". This means that the rugged landscape of real protein folding is not globally flat with totally unpredictable fluctuations as it would be for a random heteropolymer but has a preferred direction of flow. It can be described as a rugged funnel shown in Fig 4 whose shape, as we shall see, can be estimated using theory and experiment. At the bottom is a unique native state. Obvious order parameters to describe the position of an ensemble of states in the funnel are Q, the percent of native-like contacts, A, the percent of correct dihedral angles in the protein backbone, percent correct secondary structure. Other order parameters such as total helicity may also be used to classify ensemble of states: local secondary structure.

Through these order parameters, the folding landscape is stratified. Within each stratum we can define an average energy $E(Q)$, although there are still many states with different energies. To describe their distribution and properties we apply a REM model. The late stages of protein folding will have few states all highly similar to the native. These could be given specific names (if necessary) and detailed kinetics counts. These are analogous to the taxonomic substates discussed by Frauenfelder (Frauenfelder et al. (1990). As indicated in Fig.5, some routes can dead-end in a low energy misfolded conformations from which the protein has to partially unfold to reach the native state. In the early stages of folding, corresponding to a nearly denatured protein with $Q \approx 0$, there will be many states, and the ensemble language is clearly most appropriate. The hopping rate R between microstates at each stage is roughly

$$R \approx e^{-\Delta E^2(Q)/k_B T^2} \tag{4}$$

This determines the ability of the ensemble to flow between different strata. The complete statistical mechanical treatment requires knowledge of all thermodynamic variables as a function of the order parameters. In particular the functional

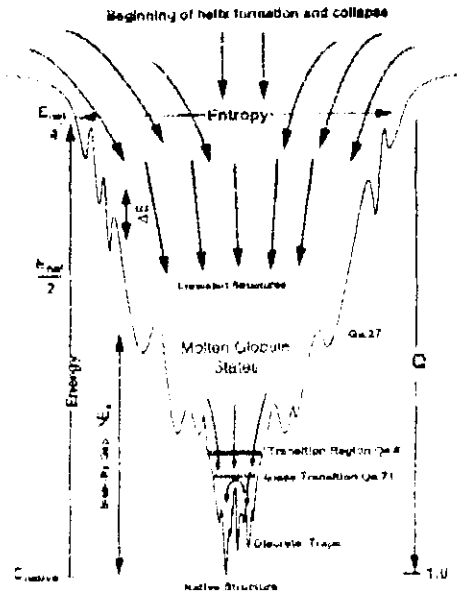

Fig. 4. Energy landscape parameters for a realistic protein folding funnel

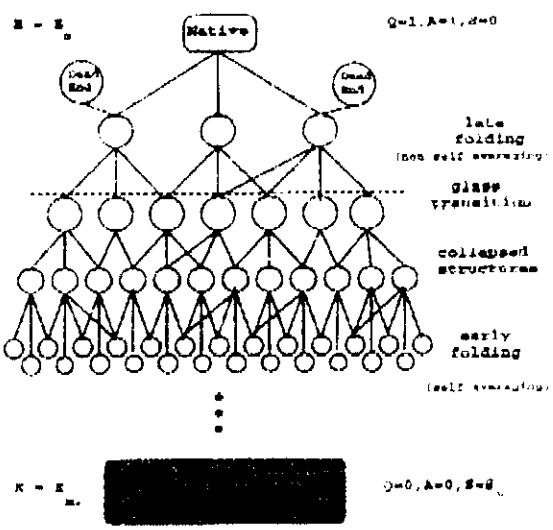

Fig. 5. Protein conformations tree structure and corresponding thermodynamic properties ordered according to various reaction coordinates describing the folding process

dependence of of the thermal average energy $\bar{E}(Q)$, the ruggedness $\sqrt{\overline{\Delta E^2(Q)}}$, the density of states $\Omega(E,q)$ or equivalently the entropy S(E,Q), and the local glass transition temperature $T_g(Q)$. Following Bryngelson and Wolynes, we derive these quantities using the simplest form of the random energy approximation in which correlations within a stratum is neglected.

According to Postulate A, the energy of a given misfolded state arises from the contributions of many random terms, so the probability distribution of energies in the interval δE at any position in the funnel is a gaussian centered about the mean energy

$$P(E)\delta E = \frac{1}{\sqrt{2\pi\overline{\Delta E^2(Q)}}} \exp\left\{ -\frac{(E-\bar{E}(Q))^2}{2\overline{\Delta E^2(Q)}} \right\} \delta E \tag{5}$$

If there are γ configurations per residue for a protein in its unfolded state, then the total number of configurations for a protein with N residues is $\Omega_o = \gamma^N$. In models using a reduced description of the protein that only include the backbone coordinates γ is less than 5 and when corrections are made for the excluded volume effect in compact configurations, $\gamma = \gamma^* \approx 1.5$ (Flory (1954) Bryngelson et al. (1995)). As the structures become more similar to the native protein, the total number of configurations will decrease since only a single backbone conformation represents the native state. If $\Omega_s(Q)$ is the number of structures with similarity measure Q to the native structure, then a rough approximation to this behavior is

$$\Omega_s(Q) = \gamma^{*N(1-Q)} \tag{6}$$

The corresponding entropy will also decrease as the native structure is approached

$$S_s(Q) = k_B log \Omega_s(Q) = k_B N(1-Q) log \gamma^* \tag{7}$$

The density of conformational states with energy E and similarity Q is then

$$\Omega(E,Q) = \Omega_s(Q) P(E) \tag{8}$$

and the total entropy

$$S(E,Q) = S_s(Q) - \frac{(E-\bar{E}(Q))^2}{2\overline{\Delta E^2(Q)}} - \log\left(\frac{\sqrt{2\pi\overline{\Delta E^2(Q)}}}{\delta E} \right) \tag{9}$$

where δE is large relative to the spacings between energy levels but small relative to $\sqrt{\overline{\Delta E^2(Q)}}$. The last term is neglected since even in the worst case it is varying logarithmically giving only a small correction. At thermal equilibrium, the most probable energy can be determined using the thermodynamic definition of the temperature

$$\frac{1}{T} = \frac{\partial S}{\partial E} \tag{10}$$

or more directly as shown in Fig 3, we can find the maximum of the thermally weighted canonical probability $p(E,Q) \propto \Omega(E,Q) e^{-E/k_B T}$,

$$E_{min}(Q) = \bar{E}(Q) - \frac{\overline{\Delta E^2(Q)}}{k_B T} \tag{11}$$

The number of thermally occupied states and entropy associated with this most probable energy are

$$\Omega(E_{m.p.}, Q) = \exp\left[\frac{S_o(Q)}{k_B} - \frac{\Delta E^2(Q)}{2(k_B T)^2}\right] \tag{12}$$

$$S(E_{m.p.}, Q) = S_o(Q) - \frac{\Delta E^2(Q)}{2k_B T^2} \tag{13}$$

Combining 11 and 13 the free-energy of the misfolded structures with configurational similarity Q and at a fixed temperature is

$$F(Q,T) = E_{m.f.}(Q) - TS(E_{m.p.}, Q)$$
$$= E(Q) - \frac{\Delta E^2(Q)}{2k_B T} + TS_o(Q) \tag{14}$$

Folding is considered to be a two-state reaction, $unfolded \rightarrow folded$ so that under some thermodynamic condition, the free energy has a double minimum. As seen in Fig 6, one minimum lies near the folded state $Q \approx 1$ and the other is at the position $Q_{min} \approx 0$ where the free energy of the misfolded states has a minimum. This minimum can either be the random coil state or a collapsed state with some degree of ordering. To a first approximation, we can neglect the entropy of the folded state so that its free energy is equal to its internal energy, E_N. At the folding temperature T_f, the probability of being in the folded state is equal to the probability of being in the misfolded state. Equating the free-energy of the folded and misfolded states at the folding temperature, $F_{native} = F(Q_{min}, T_f)$, yields expressions for the slope of the funnel stability gap $\delta E_s = E(Q_{min}) - E_N$

$$\delta E_s / T_f = S_o + \Delta E^2(Q_{min})/2k_B T_f^2 \tag{15}$$

which is related to the stability gap $\delta E_s = E(Q_{min}) - E_N$ and for the folding temperature

$$T_f = \frac{\delta E_s + \sqrt{\delta E_s^2 - 2S_o \Delta E^2(Q_{min})/k_B}}{2S_o} \tag{16}$$

Since Q_{min} is close to the unfolded state, we will consider the folding temperature as being referenced to a set of states with little structural similarity to the native state, $Q \approx 0$.

Recall a glass transition occurs at the temperature where there are too few states available so the system remains frozen in one of a few distinct states. Within each stratum this is characterized by an entropy crisis where $S(T_g, Q) = 0$. Using 13 the local glass transition temperature is

$$T_g(Q) = \sqrt{\frac{\Delta E^2(Q)}{2k_B S_o(Q)}} \tag{17}$$

Local glass transition temperatures are manifested in the folding and collapse times measured in lattice calculations which are summarized in Fig.7. Analytical and numberical studies on lattice models have shown that the ratio of T_f/T_g can

be used to distinguish fast and slow folding sequences. This ratio also plays a central role in developing energy functions to predict protein structures. Calculating this ratio using the set of states with the least structural similarity to the folded state gives

$$\frac{T_f}{T_g} = \sqrt{\Lambda + \sqrt{\Lambda - 1}} \tag{18}$$

where

$$\Lambda = \frac{k_B}{2S_o}\left(\frac{\delta E_s^2}{\Delta E^2}\right) \tag{19}$$
$$\sim \left(\frac{\delta E_s^2}{2\Delta E^2}\right)\frac{1}{N\log\gamma^{-1}}$$

For a protein to fold, Λ and consequently T_f/T_g must be greater than 1, and since S_o, ΔE^2 and E_N all depend linearly on the chain length N, T_f/T_g is independent of length and sensitive to the interaction energies.

A phase diagram is a useful tool for summarizing what states of a protein are involved in the various folding scenarios. So far our analysis has only used a single parameter Q to characterize the changes in free energy and the differences between the native and unfolded states. Clearly there are other parameters besides the number of correct contacts that could be used to compare structures and describe the partial ordering that occurs as the protein folds. For example, in a folded protein the core consists primarily of hydrophobic residues and the surface, of hydrophilic residues. This ordering is due to the hydrophobic effect arising from folding a protein in water. Variations in the solvent properties will have profound effects on the interaction energies of the hydrophobic groups. The phase diagram in Fig 6 shows the possible thermodynamic states of a protein as a function of temperature and the roughness of the energy landscape. The phase diagram is actually a slice through a more complicated diagram along a line of some average hydrophobicity of the sequence. Since average hydrophobicity itself depends on solvent and temperature under some conditions the coexistence curve between the random coil and the folded state disappears, and the folded state becomes only accessible after non-specific collapse. The thermodynamic dependence of hydrophobic forces is one of the main complicating features in relating the theoretical phase diagrams (that assume temperature independent forces) to experiment. This is most manifest in the phenomena of cold and pressure induced denaturation. Approximate analytical expressions for coexistence curves between the collapsed phase and the random coil and frozen phase were derived by Sasai and Wolynes(1990) using the associative memory Hamiltonian given at the end of this chapter to describe the interaction energies between residues. The glass transition, which occurs after the collapse of the system, is a continuous transition. This portion of the phase diagram is typical of random heteropolymers (Dinner ,et al. (1994). Recent lattice simulations of Socci and Onuchic (1995) on proteinlike sequences probe the phase diagram as a function of temperature and average hydrophobicity and provide qualitatively the same picture.

This phase diagram can be used to understand the free energy behavior in the various folding scenarios. In general the free energy is either unimodal or bimodal. Type 0A and Type I scenarios dominate the left-hand part of the phase diagram in which no glass transition occurs, and the system is at a temperature such that the global minimum is the native folded states with Q=1. Direct folding from the random coil state favors these scenarios since glassy states can only occur in nearly collapsed chains. Type II scenarios occur in the right-hand part near the coexistence curves for the folded, collapsed, and collapsed frozen states. In Type IIA the glass transition occurs after the thermodynamic barrier and in Type IIB the folding protein becomes glassy before the barrier is reached. Theoretical calculations of Takada and Wolynes ? have provided a more detailed description of the crossover between the different activated folding scenarios.

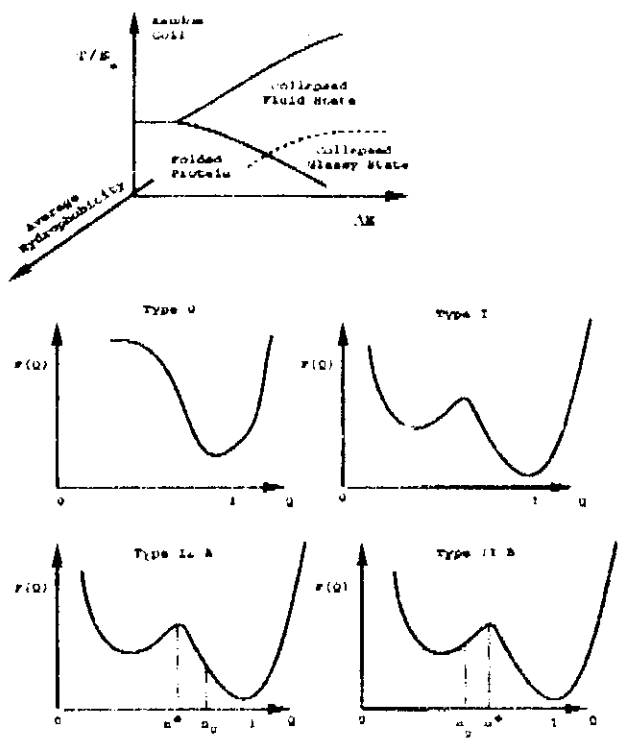

Fig. 6. Phase diagram and corresponding free energy curves for a folding protein

3 Simple Models of Folding Kinetics of MFHP

A protein folding along the funnel shown in Fig.4 moves through an ensemble of partially ordered structures characterized by the similarity measure Q. The gradient of the free energy determines the instantaneous drift velocity down the funnel. The roughness at any stage acts like a set of speed bumps slowing this drift. Superimposed on the drift are stochastic fluctuations in Q reflecting individual escapes from traps. A simple description of the overall dynamics within the folding funnel arising from these effects is obtained using a diffusion equation. At a given temperature the population of the various structural strata changes with time according to

$$\frac{\delta P(Q,t)}{\partial t} = \frac{\partial}{\partial Q}\left\{ D(Q,T)\left[\frac{\partial P(Q,t)}{\delta Q} + P(Q,t)\frac{\partial \beta F(Q,T)}{\partial Q}\right]\right\} \qquad (20)$$

In general the local configurational entropy D depends on the roughness of the energy surface which determines the escape time from traps. At sufficiently high temperatures it crudely follows a Ferry law typical of glasses

$$D(Q,T) = D_o \exp\left[-\Delta E^2(Q)/(k_B T)^2\right] \qquad (21)$$

The functional form of $D(Q,T)$ at temperatures near the glass transition temperature is a bit more complicated and depends on the nature of local moves. The above form is motivated by the following qualitative argument. The diffusion coefficient is inversely proportional to the lifetime $\tau(Q)$ of a microstate with similarity Q to the native state. If the microstate is deep, it will be long-lived and the diffusion coefficient becomes small. To avoid being trapped in this microstate characterized by a roughness $\Delta E(Q)^2$, motion must take place over an energy barrier $E(Q) - E_{min}(Q) = \Delta E^2(Q)/k_B T$ in the time τ_o it takes for a large segment of the chain to move. This gives an escape time from the local traps that is super Arrhenius

$$\tau(Q) = \tau_o \exp\left[\Delta E^2(Q)/(k_B T)^2\right] \qquad (22)$$

In the case of fast downhill folding at a fixed temperature shown in Type 0 scenario a kinetic folding bottleneck occurs at a region $Q^{\ddagger}_{kin}$ with the maximum lifetime or the smallest diffusion coefficient. This maximum lifetime is also a simple estimate of the overall folding time

$$\tau_f \approx \tau_{max}(Q^{\ddagger}_{kin}) \qquad (23)$$

For a bistable system as in Type I and Type IIA scenarios, the overall folding time τ_f will be determined by the difficulty to overcome the free energy barrier and a prefactor that depends on the ruggedness of the energy landscape

$$\tau_f \approx < \Delta Q^2_{MG} > D^{-1}(Q^{\ddagger}) e^{\Delta F^{\ddagger}/k_B T} \qquad (24)$$

where $\Delta F^{\ddagger}$ is the free energy barrier measured from the unfolded minimum to the top of the thermodynamic barrier. $< \Delta Q^2_{MG} >$ is the mean square fluctuation

of the configuration coordinate in the molten globule state This equation suggests that an Arrhenius plot of folding time versus inverse temperature would be curved, and such behavior is frequently observed in protein folding experiments (Creighton (1994)) and lattice simulations as seen in Fig.7. As the temperature is decreased, the escape time will increase until the local glass transition temperature $T_g(Q)$ is reached. For $T < T_g(Q)$, the protein has kinetic access to very few structures and the protein is effectively frozen into a single or several low energy states. In this case the kinetics are dominated by the details of the specific landscape and the expressions for the folding time and the diffusion coefficient have to be modified.

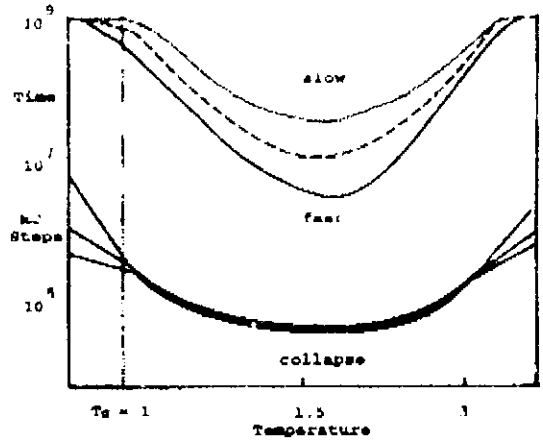

Fig. 7. Folding and collapse times of designed 27 mers

Using a simple three-dimensional lattice model and Monte Carlo dynamics, the collapse and folding times for designed sequences of proteinlike heteropolymers have been studied and a sketch of these results is given in Fig.7 (?, Bryngelson et al (1995)) In all cases, the polymers are 27 monomers long and possess a maximal compact non-degenerate native state with 28 contacts. The simulations start out with the polymer in a fully extended form, and the protein is considered folded when all 28 correct contacts and collapsed when any 25 contacts are made. The times given in the curves are the number of Monte Carlo steps required to first reach either the folded or a collapse conformation The lower curve shows that the collapse time is independent of sequence or self-averaging over a wide range of temperatures until the glass transition temperature is reached. The folding curve exhibits a much greater spread in the times. The fast folder is the sequence with the least frustration (lowest energy) and the largest T_f/T_g ratio, while the slow folder has highest energy and $T_f/T_g < 1$. The folding temperature T_f is defined in Fig.8 as the temperature where the probability of the occupancy of the native structure is one-half. At the glass transition temperature, the fast

folder is over 90% in the native state while the slow folder has a native state population less than 20%. Clearly at this low temperature the native state is thermodynamically more stable, but the system is now getting trapped in local minima so that it is no longer kinetically accessible. For a protein of significant length to be foldable on biological time scales, the folding temperature must be greater than the glass transition temperature.

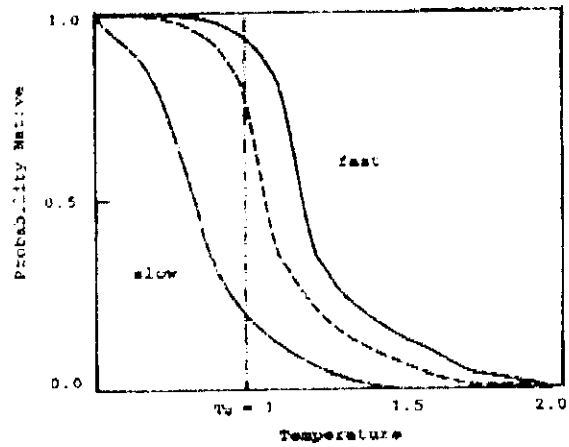

Fig. 8. Temperature dependence of the probability density of the native state for designed 27 mer

Onuchic et al. developed a law of corresponding states to relate simulations of small lattice models to real proteins. The correspondence analysis made use of a theory of helix formation in collapsed polymers that related the configurational entropy S_c to the amount of helical structure. For a 60-amino acid chain at 60% helicity, $S_c \approx 40k_B$ which corresponds approximately to a conformational entropy of $0.6k_B$ per monomer unit. Assuming reconfiguration times on a rough energy landscape are on the order of the folding time of milliseconds observed experimentally, expression 23 suggests that the roughness of the landscape at the folding temperature $\Delta E^2/2k_BT_f^2$ ranges from 11 to 18. These two quantities allow them to estimate the slope of the funnel according to 15, $\delta E_s/k_BT_f \approx 58$, the dimensionless ratio of the energy gradient to the ruggedness $(\delta E_s/k_BT_f)/(\sqrt{\Delta E^2/2k_BT_f^2}) \approx 14$ and the ratio $T_f/T_g \approx 1.6$. Comparable values for T_f/T_g are obtained for lattice simulations of 27mers with a three-letter code implying that every lattice bead corresponds to about two amino acids. The funnel in Fig.4 reflects the correspondence we have just advanced. At T_f folding proceeds via a Type IIB scenario with the transition state at Q=0.60 and a glass transition at Q=0.71. Although the funnel was originally using the correspondence with the lattice simulations, one significant aspect of its form form

was dramatically confirmed by NMR measurements of Huang and Oas (1995) on the submillisecond folding of the λ repressor. In their experiments the thermodynamic bottleneck for folding occurs when approximately half the native contacts are made, as indicated by the sensitivity of the folding rate to added denaturant.

4 Protein Structure Prediction Using Optimized Energy Functions

Nature seems to have designed its proteins to have T_f/T_g greater than 1, so how can we use this criterion to design energy functions for protein structure prediction and what are the further problems to be faced? Molecular biology gives us the sequence of a protein $\{q_i\}$ and in the cases where the protein has been well crystallized or is small enough for structure determination by NMR, the mean positions $\{r_i\}$ of all the atoms. Considerable effort has been invested in developing energy functions based on standard bonding and van der Waals interactions. These have been shown to be well parameterized to describe the motion of the atoms about or near their crystallographic coordinates. It is not yet possible to test whether they are sufficient to describe the folding process since the natural process is very slow on the time scale of atomistic simulations ($\gtrsim$ 1 μsec). Thus we seek simpler energy functions that can encode the sequence-structure correlation. We are helped by the fact that evolution has already solved the problem how to find different sequences $\{q_i\}$ compatible with a given structure $\{r_i\}$. For scientist trying to break the protein folding code the problem is to use all the information that biologists and crystallographers have collected in their sequence and structural databases, $\{q_i\}^\mu$ and $\{r_i\}^\mu$, to develop new energy functions that fold these proteins and can be generalized to fold the plethora of new sequences arising from the genome project. Code breakers who wish to take this phenomenological route must understand both evolution and physics.

Friedrichs and Wolynes introduced in 1989 an associative memory Hamiltonian (H_{AMH}) that encodes correlations between the sequence of the target protein whose structure is to be determined and the sequences and structures of a set of memory proteins taken from the database. The associative memory Hamiltonian resembles the empirical energy functions used in conventional molecular dynamics, but its form was motivated by energy functions used in neural network theory to perform pattern recognition (Goldstein et al. (1992)).

$$\mathcal{H}_{AM}(\{r_i\}) = \sum_{\mu} \sum_{i<j} \gamma_{ij}^{\mu} \theta(r_{ij} - r_{ij}^{\mu}) + \mathcal{H}_c \tag{25}$$

In its simplest form, the associative memory Hamiltonian is a function of the pairwise distance between the α-carbons of residues i and j. r_{ij}. γ_{ij}^{μ} encodes a degree of similarity between residues i and j of the target protein and and a corresponding pair in the memory protein μ, and it may include information about the physicochemical properties of the residues, their probability of mutation, or context of the residues in the protein. $\theta(r_{ij} - r_{ij}^{\mu})$ is a Gaussian function of the

difference between the pairwise distance in the target structure and the memory structure. The energy function has a minimum of varying depth for the structure of each memory protein. These minima are differentiated by the sequence "property" similarity weights γ^μ. $\mathcal{H}_o$ is a typical chain molecule Hamiltonian for the backbone atoms and includes harmonic terms to induce backbone rigidity and the correct chirality, and to prevent the overlap of non-bonded α-carbons (Friedrichs and Wolynes (1990), Friedrichs et al. (1991)). In essence each memory protein constructs a small folding funnel. If these funnels add coherently or only a single one dominates, the Hamiltonian will not be very frustrated and a single folding funnel to a structure consistent with empirical correlations will be formed. Achieving this requires a good choice of the γ-parameters and training proteins. The quantitative version of MFP helps us do this.

The energy parameters in the energy landscape analysis can be expressed in terms of the γ coefficients in the energy function

$$\delta E_g = \mathbf{A}\gamma \tag{26}$$

$$\Delta E^2 = \gamma \mathbf{B} \gamma \tag{27}$$

The explicit values of $\mathbf{A}$ and $\mathbf{B}$ are obtained from a set of training proteins with known structures and simulated molten globule states. According to eq.(16), T_f/T_g is maximized when $\delta E_g/\sqrt{\Delta E^2}$ is maximized. This maximization procedure leads to optimal values $\gamma = \mathbf{B}^{-1}\mathbf{A}$. Simulations with naive assignments of gamma, e.g. interactions energies between similar residues being $E_{sim} = -3$ and between dissimlar residues $E_{dis} = -1$, give rise to much smaller T_f/T_g values.

Simulated annealing for the optimally-encoded Hamiltonian generally leads to qualitatively correct structures when the target protein can be assigned to one of three broad classes of folding motifs: alpha, beta, or mixed alpha-beta proteins (Goldstein et al. (1992)). For a comparision code based only on hydrophobicity and proximity, the results of such a molecular dynamics run for the myoglobin shown earlier are presented in Fig 9. The simulation begins with the protein in a random extended form with a radius of gyration typical of a random coil, $R_g = 60\mathring{A}$. Collapse and compaction of the protein occurs quickly to a state that has roughly the correct topology or fold, but has incomplete secondary structure. The local Q score for helix-A is about 0.3 when the collapsed protein is first formed. Continued folding in this compact state completes the formation of the helices and modifies the tertiary contacts. In the short trajectory shown here roughly one-third of all the native contacts have been formed, but the local Q value is considerably higher. Preliminary studies of the local T_f/T_g ratios indicate that helix-A is the major part of a kinetically competent, quasi-independent folding unit or foldon (Panchenko et al. (1996)). The associative memory Hamiltonian can function as a laboratory for similar studies with more realistic encodings and provides a powerful tool to determine the energy landscape parameters at various stages in the folding funnel.

The energy landscape perspective presented here is a simple, but powerful framework to view the complex nature of protein folding. Experimental techniques are just being developed to study the submillisecond stages of folding

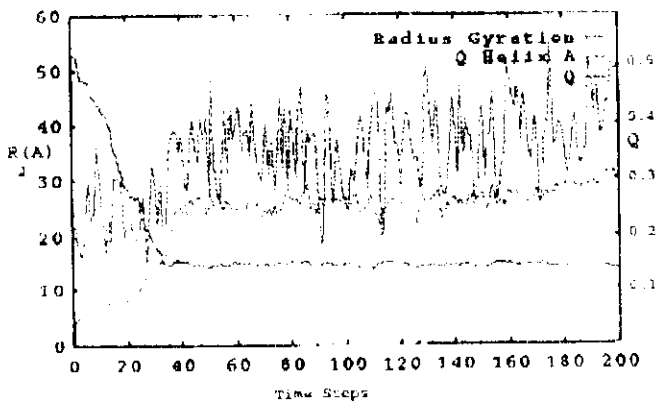

Fig. 9. Molecular dynamics simulation of the collapse and structure formation in myoglobin using the H_{AMH} energy function

in which the protein is compact and partially folded. This region and the late stages of folding present challenges to statistical physicists because the partially folded protein is in a low entropy state. This low entropy density state appears to be liquid crystalline like in which subtle forces can cause subtle forms of partial ordering. The diverse routes leading to the native state of a protein can be better probed by adding at least one other reaction coordinate to the folding funnel in Fig.4. For example, the roles of substructures, "foldons" which independently order, and microphase separation in the folding process are of great interest. Many thermodynamic properties of proteins in these last stages of folding are non-self-averaging and can be important in determining whether a given sequence is foldable. The statistical physics view of the protein folding problem has already lead to partial successes in solving such practical problems in molecular biology as structure prediction and design. The energy landscape approach offers experimentalists a framework to analyze their folding experiments and guide the development of more precise probes of Nature's information highway.

References

Bryngelson, J., Wolynes, P. G. (1987): Spin glasses and the statistical mechanics of protein folding. Proc. Nat. Acad. Sci., U.S.A.**84**, 7524-7528

Bryngelson, J., Onuchic, J. N., Socci, N., Wolynes, P. G. (1995): Funnels, pathways and the energy landscape of protein folding: a synthesis. Proteins **21**, 167–195

Creighton, T. E. (1994): *In: Mechanisms of Protein Folding* (Oxford Univ., Oxford, Pain, R. H. editor), 1–25

Dill, K. A. et al. (1995): Principles of protein folding - A perspective from simple exact models. Protein Science **4**, 561–602

Dinner, A., Sali, A., Karplus, M., Shakhnovich, E. (1994): Phase Diagram of a model protein derived by exhaustive enumeration of the conformations. J. Chem. Phys **101**, 1444-1451

Flory, P. J. (1954): *Principles of Polymer Chemistry* (Cornell Univ., Ithaca). 523–530

Frauenfelder, H. et al. (1990): Proteins and pressure. J. Phys. Chem. **94**, 1024–37

Friedrichs, M., Wolynes, P. G. (1989): Toward protein tertiary structure recognition by means of associative memory hamiltonians. Science **246**, 371–373

Friedrichs, M., Wolynes, P. G. (1990): Molecular Dynamics of Associative Memory Hamiltonians for Protein Tertiary Structure Recognition. Tetrahedron Comp. Method. **3**, 175

Friedrichs, M., Goldstein, R., Wolynes, P. G. (1991): Generalized protein tertiary structure recognition using associative memory hamiltonians. J. Mol. Biol. **222**, 1013–1034

Garel, T., Orland, H., Thirumalai, D. (1995): *In: New Developments in Theoretical Studies of Proteins: Advanced Studies in Physical Chemistry* (Ron Elber, editor, World Scientific, Singapore)

Gierasch, L. M., King, J. (1990): *Protein Folding:Deciphering the Second Half of the Genetic Code* (AAAS, Washington) vii–vii

Goldstein, R., Luthey-Schulten, Z. Wolynes, P. G. (1992): Optimal protein-folding codes from spin-glass theory. Proc. Natl. Acad. USA **89**, 4918-4922

Huang, G. S., Oas, T. G. (1995): Proc. Natl. Acad. Sci. **92** 6878

Luthey-Schulten, Z., Ramirez, B., Wolynes, P. (1994): Helix-coil, liquid-crystal and spin-glass transitions of a collapsed heteropolymer. J. Phys. Chem. **98** 2177–2185

Onuchic, J. N., Wolynes, P. G., Luthey-Schulten, Z., Socci, N. D. (1995): Toward an outline of the topography of a realistic protein-folding funnel. Proc. Nat. Acad. Sci., U.S.A.**92**, 3626–3630

Panchenko, A., Luthey-Schulten, Z. Wolynes, P. G. (1996): Foldons, Protein Structural Modules and Exons. Proc. Nat. Acad. Sci., U.S.A. to be published

Sasai, M., Wolynes, P. G. (1990): Molecular theory of associative memory hamiltonian models of protein folding. Phys. Rev. Lett. **65**, 2740–2743

Shakhnovich, E., Gutin, A. (1990): Implications of thermodynamics of protein folding for evolution of primary sequences. Nature **346**, 773–775

Socci, N., Onuchic, J. (1994): Folding kinetics of protein-like heteropolymers. J. Chem. Phys. **101**, 1519–15[illegible]

Socci, N., Onuchic, J. (1995): Kinetic and Thermodynamic Analysis of proteinlike heteropolmers: Monte Carlo histogram technique. J. Chem. Phys **103**, 4732–4744