Protein Folding Modeled through Spin Glasses

Dominic A. Ricci
Dept. of Physics, University of Illinois at Urbana-Champaign

May 2, 2000

Abstract

The concept of protein folding and the interactions which cause it are introduced. The characteristics of protein folding are compared to those of both glass and spin glass systems. The Hamiltonian for modeling protein folding as a spin glass is stated and then converted to a stochastic form in the manner of Bryngelson and Wolynes. The equilibrium thermodynamics and conformational dynamics of the system are explored, with special emphasis on the critical temperature and behavior. The stochastic spin glass model is used to estimate the average folding time of a protein. The validity of the model is assessed and critiqued.

1 Introduction

Proteins are complex organic polymers composed of amino acids connected by peptide bonds. Because every amino acid contains a functional group, there exists interactions between the various monomers. These interactions cause the linear protein chains to assume secondary and tertiary structures through a process called folding. Although secondary structures such as α helices and β sheets are formed through simple hydrogen bonding, the folding that results in the creation of tertiary structures is caused by a more complex array of interactions. This tertiary structure is of interest to researchers because it determines both the functionality and mechanisms of activity of a given protein. A conformational modification to the tertiary structure of a protein significantly alters its functional behavior. This paper will focus on the folding associated with a protein assuming its native, or fully folded, tertiary form.

1.1 Protein Structure Compared with Glasses

As shown through x-ray diffraction studies, proteins have definite structures. There exists evidence, however, that proteins retain an amount of residual flexibility in their native folded states. Rather than occupying only unique fixed positions, proteins exist in a larger portion of their allowed conformational space, and thus there are many conformational substates. Another characteristic of proteins is that they are composed of both hydrophilic and hydrophobic surfaces. A protein chain in solution will fold such that it maximizes its interior hydrophobic and its external hydrophilic surface areas, thus forming a tightly packed globular structure.

The fact that proteins exhibit both close packing behavior as well as conformational flexibility led to their comparison with glasses, and it was for this reason that Schrodinger described them as "aperiodic crystals" in "What is Life?". Glasses are amorphous substances that are not at equalibrium and that have no minimum in free energy. This analogy, however, is not entirely appropriate as the observed behavior of proteins does not reflect those characteristics of glasses. The process of melting and cooling a glass produces a final microstructure which is different from its initial structure. The similar process of annealing a given protein results in a unique distribution of conformations. Additionally, the thermal fluctuations of a protein do not disturb its tertiary structure on account of this uniqueness. This characteristic is in direct opposition to the non-equalibrium behavior of glasses. A more appropriate model for protein behavior is that of the spin glass [3].

1.2 Proteins and Spin Glasses

There are several experimentally observed characteristics of protein folding that are relevant to the spin glass model. While a protein's native state is well

defined, multiple distinguishable conformations can correspond to that state. This suggests that the native state of a protein is actually composed of multiple conformations at the same energy. In spite of the complexity of the conformational pathways, protein folding appears to occur in a few steps and at a relatively fast rate. This second behavior runs contrary to Leventhal's paradox, a statistical argument that the direct walk folding time for a single polypeptide is longer than the age of the universe. If a protein with one hundred residues had ten possible conformational states, then there are 10^{100} possible states of the chain. Stepping through conformational space at 10^{-13} seconds per step, it would take 10^{77} years for one protein to attain its native state. This dilemma is resolved if there is more than one conformational state corresponding to the native state of the protein [3].

A spin glass refers to a system of "quenched, random, frustrated interactions among a large set of nearly identical, simple objects". A frustrated system is one in which there is no trivial single configuration that will satisfy all interactions simultaneously (see fig. 1).

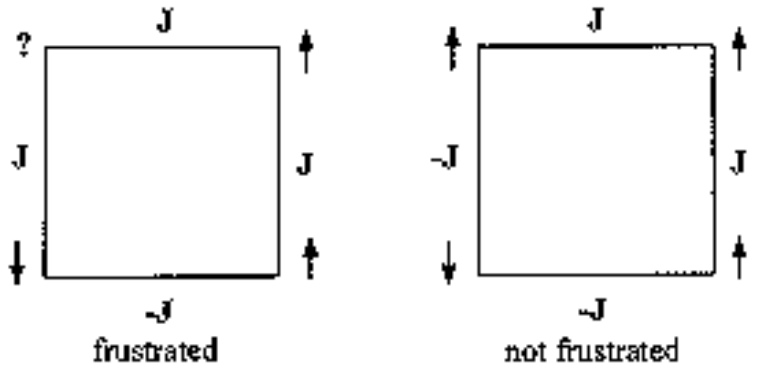

Figure 1: The idea of frustration on an elementary plaquette. Nearest neighbor Ising model with random interactions. Four spins with interactions that are either: $+J$ (ferromagnetic) or $-J$ (antiferromagnetic) [3].

Rather than settling into a single ground state as it cools, a spin glass system will exist in a multitude of lowest energy states. On a large scale such a system has a degenerate frustrated ground state. This differs from normal glasses for which the assumption is made that there exists a non-degenerate ground state. Through their degenerate ground states spin glasses reflect the multitude of conformational states exhibited by the native states of folding proteins, thus suggesting that protein folding should be modeled as a spin glass system [3].

2 The Bryngelson/Wolynes Spin Glass Model of Protein Folding

A protein can be pictured through a coarse-grained model where the overall conformation of the protein is described through the sum of the discrete conformational states of each individual component amino acid. Every amino acid has one native conformation and ν other non-native conformations. Additionally, it should be noted that the systematic frustration is lowest in the native state of a protein, which is the principle of minimal frustration.

2.1 The Protein Folding Hamiltonian

The Hamiltonian for this system will be the sum of the interactions involved in protein folding. First, the energy associated with the state of each amino acid is $-\epsilon_i(\alpha_i)$, where i indexes the amino acid and α_i refers to the conformational state of the ith amino acid. A second set of interactions are those which occur along the polypeptide chain, such as hydrogen bonding in α-helices. These interactions are standardly approximated as occurring only between nearest neighbors. The energy of each of these interactions is $-J_{i,i+1}(\alpha_i, \alpha_{i+1})$. A final set of interactions are long-range ones that occur when bends in the peptide chain bring two separated amino acids into proximity. The energy of these interactions are denoted by $-K_{i,j}(\alpha_i, \alpha_j, r_i, r_j)$, where r_i is the position of the ith amino acid. A simple interpretation of these terms relates $\epsilon_i(\alpha_i)$ to the primary structure of the protein, $J_{i,i+1}(\alpha_i, \alpha_{i+1})$ to the secondary structure, and $K_{i,j}(\alpha_i, \alpha_j, r_i, r_j)$ to the tertiary structure. Applying this model, the Hamiltonian describing the protein is

$$E = -\sum_i \epsilon_i(\alpha_i) - \sum_i J_{i,i+1}(\alpha_i, \alpha_{i+1}) - \sum_{i,j} K_{i,j}(\alpha_i, \alpha_j, r_i, r_j). \qquad (1)$$

As a matter of notation, the index for the native state of an amino acid is 0, with the non-native states as the integers from 1 to ν, where ν is on the order of 10. This Hamiltonian will be examined for a protein composed of N amino acids of which N_0 are in their native state [1].

2.2 The Stochastic Approximation

Because the above Hamiltonian is a complicated function of α_i and r_i, the brute force application of statistical mechanics to this problem does not produce an easily obtainable solution. Bryngelson and Wolynes produceda more tractable model by examining the energy distributions associated with different microstates. Let the energy of the protein molecule be a random variable

distributed similarly to the energy levels of the Hamiltonian. While a conventional spin glass system has its spins fixed on a lattice, the protein sequence is free to move in space, and thus the assumption is made that the energies of different protein conformations are uncorrelated. From this random-energy approximation, it can be seen that the joint probability distribution for n configurations with N_0 amino acids in their native states and energies E_i is given by $P(E_1, ..., E_n, N_0) = \Pi_{i=1}^{n} P(E_i, N_0)$.

The Hamiltonian will be simplified if its energy terms are stoichastically distributed. The primary structure energy term can be distributed such that the native state has an energy of $-\epsilon_0$, while the non-native states are distributed such that they have a mean energy $-\bar{\epsilon}$ ($\epsilon_0 > \bar{\epsilon}$) and a standard deviation $\Delta\epsilon$. Similarly, the native state secondary structure energy is $-J$ ($J > \bar{J}$) and the energy of the non-native state secondary structure is distributed around a mean $-\bar{J}$ with a standard deviation of ΔJ. The tertiary structure energies are distributed with similarly defined values $-K$,$-\bar{K}$, and ΔK. Note that for all three terms, the energy of the native state is lower than the mean energy of the non-natives states, as to be in agreement with the principle of minimal frustration. In the case of the tertiary structure interactions, the forces are short range and thus only significant for adjoining amino acids. This number, z, is dependent on the degree of folding and is usually on the order of 2 to 3.

The final simplifying assumption is that the native and non-native amino acids are randomly distributed within the protein. As such, the distribution of energies following the stochastization of the Hamiltonian is given by the Gaussian probability distribution $P(E, N_0)$ with a mean energy of

$$\bar{E}(N_0) = -N_0\epsilon_0 - \frac{N_0^2}{N}L - (N - N_0)\bar{\epsilon} - (N - \frac{N_0^2}{N})\bar{L} \tag{2}$$

where $L = J + zK$ and $\bar{L} = \bar{J} + z\bar{K}$, and a standard deviation of

$$\Delta E(N_0) = [(N - N_0)\Delta\epsilon^2 + (N - N_0^2/N)\Delta L^2]^{\frac{1}{2}} \tag{3}$$

where $\Delta L^2 = \Delta J^2 + z\Delta K^2$ [1].

The thermodynamics of the protein folding system can be calculated using this statistical model of the energy distribution.

2.3 Protein Folding Thermodynamics

In order to determine if the protein folding undergoes a phase transition, it is necessary to calculate the free energy of the system. This can be accomplished by determining the entropy of the system and then applying the identity $\frac{1}{T} = \frac{\partial S}{\partial E}$. The entropy of the system can be determined by examining the average density of energy levels

$$\bar{n}(E) = \sum_{N_0=0}^{N} \frac{N!}{N_0!(N - N_0)!} \nu^{N-N_0} [2\pi\Delta E(N_0)^2]^{-\frac{1}{2}} e^{-\frac{[E-\bar{E}(N_0)]^2}{2\Delta E(N_0)^2}} \tag{4}$$

in the thermodynamic limit ($N \to \infty$). In this limit the sum is equal to its maximum term, so the entropy is

$$\frac{S(E,\rho)}{N} = -\rho \ln \rho - (1-\rho)\ln(1-\rho)\ln(\frac{1-\rho}{\nu}) - \frac{[\epsilon + \bar{\epsilon} + \bar{L} + (\epsilon_0 - \bar{\epsilon})\rho + (L - \bar{L})\rho^2]^2}{2[\Delta\epsilon^2 + \Delta L^2 - \Delta\epsilon^2\rho - \Delta L^2\rho]} \tag{5}$$

where $\rho = \frac{N_0}{N}$. Taking the derivative with respect to energy obtains the energy per amino acid as a function of temperature

$$\epsilon(T) = -(\bar{\epsilon} + \bar{L}) - \frac{(\Delta\epsilon^2 + \Delta L^2)}{T} - (\epsilon_0 - \bar{\epsilon} - \frac{\Delta\epsilon^2}{T})\rho - (L - \bar{L} - \frac{\Delta L^2}{T})\rho^2. \tag{6}$$

Substitution of this equation into eqn. 5 produces the entropy per amino acid as a function of temperature

$$s(T) = -\rho \ln \rho - (1-\rho)\ln(\frac{1-\rho}{\nu}) - \frac{\Delta\epsilon^2 + \Delta L^2 - \Delta\epsilon^2\rho - \Delta L^2\rho^2}{2T^2}. \tag{7}$$

The free energy per amino acid, $f = \epsilon - Ts$, is

$$f(T) = -\bar{\epsilon} - \bar{L} - \frac{\Delta\epsilon^2 + \Delta L^2}{2T} - (\epsilon - \bar{\epsilon} - \frac{\Delta\epsilon^2}{2T})\rho - (L - \bar{L} - \frac{\Delta L^2}{2T})\rho^2 + T\rho\ln\rho + T(1-\rho)\ln(\frac{1-\rho}{\nu}). \tag{8}$$

A plot of the free energy as a function of ρ is included in fig. 2. It can be seen that a critical temperature exists that satisfies the condition $\frac{\partial f}{\partial T} = 0$,

$$T_c = \left\{ \frac{\Delta\epsilon^2 + \Delta L^2 - \Delta\epsilon^2\rho - \Delta L^2\rho^2}{2[-\rho\ln\rho - (1-\rho)\ln(\frac{1-\rho}{\nu})]} \right\}^{\frac{1}{2}}. \tag{9}$$

The system undergoes a phase transition at T_c, as the entropy of the system decreases to and then remains at zero. This occurs because the protein has become "frozen" in one of its degenerate local minima low energy conformations. For $T < T_c$, the system is frozen in this glass-like phase. In this frozen phase each protein in an ensemble would have a definite, though not identical, conformation.

The Bryngelson/Wolynes model has four phases: an unfolded phase where the proteins are in non-native conformations; a folded phase where the proteins are in their native conformation; a frozen phase where the native structure is favored; and a misfolded phase where the non-native structure is favored. This is depicted in the phase diagram in fig. 3. The most notable ramification of the thermodynamic calculations on protein folding is the first order phase transition to the frozen phase, as the numerous misfolded conformations in this phase produces relatively slow dynamics [1].

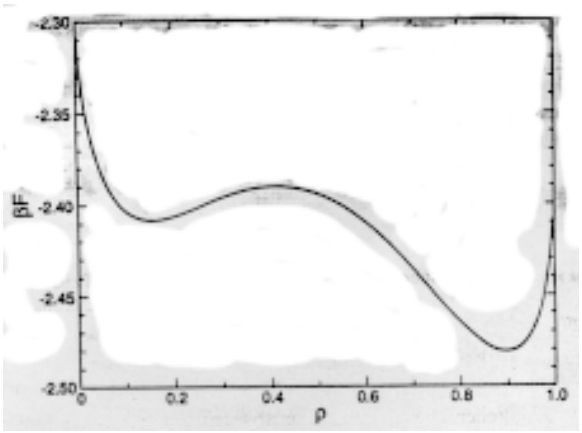

Figure 2: The free energy function for the Bryngelson/Wolynes model with $\epsilon - (\Delta\epsilon^2/2T) =$ -0.2T and $L - (\Delta L^2/2T) = 2.6T$. Note that this is the free energy, not the free energy per amino acid [2].

3 Conformational Dynamics

3.1 Transition Rates and the Critical Temperature

The critical temperature for the transition to the frozen phase can also be examined in the context of the movement of the protein between conformations. This is done by determining the transition rates associated with folding out of the glass-like low energy states. In the simplest dynamical interpretation of our model, one amino acid changes state at a time. A single transformation takes the protein into any of $N\nu$ other conformations. For a protein that originally has N_0 native state amino acids, there are $N(1-\rho)$ ways to change the conformation to add a native amino acid, $N\rho\nu$ ways to subtract a native amino acid, and $N(1-\rho)(\nu-1)$ ways to maintain the same number of native amino acids.

Letting $\bar{\epsilon} = \bar{J} = \bar{K} = 0$, the probability of a protein with N_0 native amino acids having an energy between E and $E + dE$ is

$$g(E,\rho)dE = [2\pi\Delta E(\rho)^2]^{-\frac{1}{2}} e^{-\frac{[E-\bar{E}(\rho)]^2}{2\Delta E(\rho)^2}} dE \tag{10}$$

where $\bar{E}(\rho) = N[-\epsilon\rho - L\rho^2]$ and $\Delta E(\rho)^2 = N[\Delta\epsilon^2(1-\rho) + \Delta L^2(1-\rho)]$. Another factor in the dynamic picture of protein folding is the transfer rate between states. The assumption is made that the rate of a protein moving from state A to state B is $R = R_0 e^{[-(E_B - E_A)/T]}$ for $E_B > E_A$ and $R = R_0$ for $E_B < E_A$.

A state which is lower in energy than all the adjacent conformations is a local energy minimum. The probability that a conformation is a local minimum is

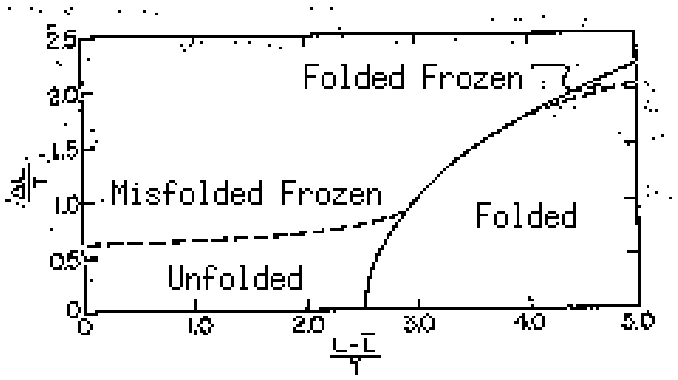

Figure 3: The calculated phase diagram of the Bryngelson/Wolynes model for $\epsilon_0 - \bar{\epsilon} = \Delta\epsilon = 2.18T$. First order phase boundaries are indicated by solid lines, and second-order phase boundaries are indicated by dashed lines [1].

$$P_{LM}(E_0, \rho) = \left[\int_{E_0}^{\infty} dE g(E, \rho)\right]^{N\nu}. \tag{11}$$

The rate of the protein transitioning out of a local minimum conformation is the sum of the rates of the protein transitioning from the the local minimum state to each adjacent conformation,

$$R = R_0 \sum_i e^{-\frac{E_i - E_0}{T}}. \tag{12}$$

Following the calculations of Bryngelson and Wolynes, the probability P_{LM} can be rewritten as a function of the transition rate

$$P_{LM}(R, \rho) = \left(\frac{1}{2\pi}\right)^{\frac{1}{2}} \left(\frac{1}{R_0}\right) \left(\frac{T}{\Delta E(\rho)}\right) e^{-\frac{T^2 \ln^2[\frac{R}{R_{sep}(\rho)}]}{2\Delta E(\rho)^2}} \tag{13}$$

for $R_{sep}(\rho) > R > R_{slow}(\rho)$,

$$P_{LM}(R, \rho) = \left(\frac{1}{2\pi}\right)^{\frac{1}{2}} \left(\frac{1}{R_0}\right) \frac{1}{[2\ln(\frac{R_0 N\nu}{R})]^{\frac{1}{2}}} \tag{14}$$

for $R_{fast}(\rho) > R > R_{sep}(\rho)$, and $P_{LM}(R, \rho) = 0$ for $R > R_{fast}(\rho)$ or $R < R_{slow}(\rho)$, where

$$R_{fast}(\rho) = \bar{R}_{LM}(\bar{E}(\rho) - (2\ln N\nu)^{\frac{1}{2}} \Delta E(\rho) \tag{15}$$

$$R_{slow}(\rho) = \bar{R}_{LM}(\bar{E}(\rho) - \{2N[-\rho\ln\rho - (1-\rho)\ln((1-\rho)/\nu)]\}^{\frac{1}{2}} \Delta E(\rho) \tag{16}$$

$$R_{sep}(\rho) = \bar{R}_0 N \nu e^{-\frac{\Delta E(\rho)^2}{2T^2}}. \tag{17}$$

From this it can be seen that there are changes in behavior at the temperatures

$$T_{high}(\rho) = \frac{\Delta E(\rho)}{[2 \ln N\nu]^{\frac{1}{2}}} \tag{18}$$

and

$$T_g(\rho) = \frac{\Delta E(\rho)}{[2N[-\rho \ln \rho - (1-\rho)\ln((1-\rho)/\nu)]]^{\frac{1}{2}}}. \tag{19}$$

Note that T_g is equivalent to the critical temperature T_c calculated in eqn. 9. Examining the temperature dependence of the transition rates reveals the behavior below the critical temperature. When $T < T_g(\rho)$, $R_{fast}(\rho) > R_{slow}(\rho) > R_{sep}(\rho)$. In this glassy phase, the distribution of transition rates out of the local minima is given by eqn. 14 and is relatively flat over the entire spectrum of rates. Very slow escape rates are common at temperatures below T_c, and this correlates well with the concept of it being a "frozen" phase with slow dynamics [2].

3.2 Folding Time

In addition to predicting the existence of the glass-like "frozen" phase, the Bryngelson/Wolynes spin glass model can also be used to investigate the time period over which protein folding occurs. By employing tools such as the continuous time random walk and the stochastic master equation, Bryngelson and Wolynes were able to combine the analysis of rate transitions with calculated thermodynamics (not included here) to calculate an expression for the average folding time, $\bar{t}$, for a protein polyer of N amino acids of which N_0 begin in their native states.

$$\ln(R_0 \bar{t}) = \bar{F}_{max} - F(\rho_b) \tag{20}$$

where $\bar{F}_{max} = max[\bar{F}(\rho)]$. This solution assumes a free energy which possesses a double well form as a function of ρ (see fig. 2, in which case ρ_b is the smaller value of ρ which corresponds to a local minimum of the free energy. A plot of $\ln(R_0\bar{t})$ is shown in fig. 4.

The parameter R_0 has been estimated to have a value between 10^8 and 10^{12} s^{-1}, so let $R_0 = 10^9$ s^{-1}. Letting $\Delta\epsilon = \Delta L = 0.30$, then a protein with 100 amino acids would fold in about 10^{-2} s, a reasonible folding time. Note how this compares favorably to the magnitude of time predicted by simple probability theory, and thus presents itself as a solution to Leventhal's paradox [2].

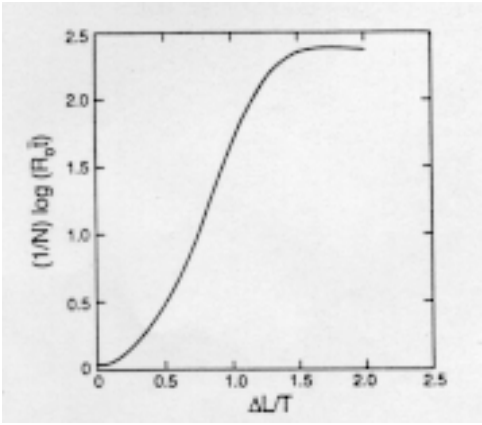

Figure 4: The folding time for the Bryngelson/Wolynes model with $\epsilon - (\Delta\epsilon^2/2T) = -0.2T$, $L - (\Delta L^2/2T) = 2.6T$, and $\Delta\epsilon = \Delta L$ [2].

4 Validity of the Spin Glass Model

Modeling protein folding as a spin glass allows for the determination of the phase behavior of the system, but it is not universally applicable. Several assumptions were made during the course of developing the model which prevent it from always being valid.

Wolynes and Bryngelson suggest that the most important deviation from real proteins in their calculations was the omission of pair correlations. The spin glass model as previously portrayed demonstrates a sharp first-order transition to the frozen phase as the temperature decreases past the critical tmperature T_c. If pair correlations, surface correlations, and collective movements are considered in the model, the effect is to produce a gradual "freezing" of the proteins. This model agrees better with experimentally observed behavior than the simpler non-correlated model [1].

Several assumptions were made in order to form a Hamiltonian which was simple enough to be converted into a stochastic form. The energy of each type of interaction (primary, secondary, and tertiary-related) is treated the same (ϵ,J,K) regardless of the specific amino acids involved. This process ignores the inherent heterogeneity of the protein sequence [1]. Additionally, the assumption that transitions occur as if one amino acid at a time is only valid if the number of amino acids that change simultaneously is much less than the total number of amino acids N. If this assumption is false, then the transition rate out of the local minimum state can no longer be considered to be of the same order of magnitude as the escape rate from the basin of attraction of that state because the "basin of attraction" is no longer well defined [**?**].

Finally, the very process of converting the Hamiltonian to a stochastic form results in a loss of accuracy. This procedure is similar to applying mean field

theory to the system, and it has been shown that while MFT produces a solvable problem, the solution is no longer exact nor is the critical behavior in agreement with experiment [4].

5 Conclusion

The problem of protein folding is of interest because proteins only become functional when they attain the proper tertiary conformation. Bryngelson and Wolynes have shown that the spin glass model is effective in predicting the critical behavior of a protein folding system. The Hamiltonian for this model accounts for the interactions which form the primary, secondary, and tertiary structures of a protein chain. This Hamiltonian, however, must be simplified through a conversion to a stochastic form in order for reasonable calculations to be performed. The thermodynamics of the spin glass model accurately determines the existence of a "frozen" glass-like phase which occurs due to a multitude of degenerate low energy local minima in conformational space. An examination of the dynamical transition rates associated with the stochastic Hamiltonian also yield the same critical behavior and temperature. Additionally, an average folding time for a protein has been calculated by Bryngelson and Wolynes from this basis that avoids Leventhal's paradox. Although the spin glass model is limited in applicability, it produces results which are more physical and in better agreement with observed behavior than simple polymer statistics.

References

[1] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987)

[2] J. D. Bryngelson and P. G. Wolynes, J. Phys. Chem. **93**, 6902 (1989)

[3] R. Clegg, class notes for PHYS 450 (Fall 1999)

[4] N. D. Goldenfeld *Lectures on Phase Transitions and the Renormalization Group* (Perseus Books, Massachusetts, 1992)