Statistical mechanics of proteins with "evolutionary selected" sequences

Sharad Ramanathan and Eugene Shakhnovich
Harvard University, Department of Chemistry, 12 Oxford Street, Cambridge, Massachusetts 02138
(Received 18 February 1994)

The requirement that the native structure of a protein be stable and kinetically accessible implies that it should correspond to a pronounced energy minimum. Thus we expect the protein sequence not to be random but selected such that this is satisfied. This is achieved in our model by defining a "selective temperature" in sequence space and statistically optimizing the sequence for the target conformation. Mean-field replica calculations are presented for this model and the phase diagram indicating the temperatures and selective temperatures at which the transition to the native conformation occurs is obtained. The transition to the native state is shown to be a first-order one. A temperature range exists in which the target structure of selected sequences is stable and kinetically accessible. It is shown that optimization at very low selective temperature leads to sequences with long-range correlations which appear to be less capable of folding.

PACS number(s): 61.44.+p, 87.15.Da, 64.60.Cn, 64.60.Kw

I. INTRODUCTION

The statistical-mechanical approach to protein folding is based on the investigation of the properties of simple models of heteropolymers. As is well known, proteins are made up of 20 different kinds of amino acids and thus the interaction energy between the different monomers of the heteropolymer can be approximated to have a Gaussian distribution (independent interaction model). The properties of such a model have been extensively studied [1–4] and it was found that there exists a transition temperature T_c (nonvanishing in the thermodynamic limit) such that below T_c the chain freezes into a small number of definite folds.

The independent interaction model corresponds to an infinite number of monomers. This model has, however, the caveat that the *interaction energies* are considered independent there while in proteins they are characterized by the *sequence* of their monomers. This motivates the study of another simplified model of proteins in which the heteropolymer is made up of only two kinds of monomers ("letters"). Such a molecule is called a copolymer. The statistical mechanics of random copolymers has been studied in the recent past [5–9].

In this "two letter" model the similar kinds of monomers attract each other and unlike monomers repel each other. This can be related to the physical picture where we classify the different monomers of the protein as either hydrophilic or hydrophobic. Though it may seem that such a system would just separate into hydrophobic and hydrophilic rich regions at low temperatures, this is not the case because of the presence of the constraints of chain connectivity. Due to the polymeric effect the position of a monomer is not independent of that of its neighbors and this leads to frustrations.

Solving this model in the mean field using the replica symmetry breaking ansatz [8] one finds that the energy levels show a continuous spectrum for large values for the free energy and a discrete spectrum for lower values. Thus as the system is cooled it "freezes" into the lower part of the energy spectrum. This is analogous to the basic physics of the independent interaction model where the same transition has been predicted.

It was shown in [3,10,8] that the random energy model (REM) [11] is a good approximation for energy surface of random heteropolymers. A number of studies concerning the dynamics of the REM [12–14] have suggested that the dynamics of freezing will be extremely slow so that the ground state, even if nondegenerate (unique), will not be reachable kinetically. This assertion was confirmed in a recent numeric study [15] where 200 random sequences were subjected to folding simulations in the model where all compact conformations were enumerated, and the ground state was known. It was found that only small fractions of random sequences were able to find their ground state conformation. Careful analysis revealed that the feature which distinguishes folding sequences is that they have a large gap in their energy spectrum, i.e., that the energy difference between the ground state and other conformations is greater in folding sequences than in nonfolding ones.

In order for a heteropolymer to be able to fold to a kinetically accessible unique native state it thus seems necessary to pull down the energy of this state far below the discrete part of its energy spectrum.

Several phenomenological models have been motivated by this idea of nonrandomness in proteins [16,17,1]. In [1] this idea was encapsulated in "the principle of minimal frustrations." The models in [16,1] assumed some special interactions between those monomers of a protein which are neighbors in its native structure. These interactions were responsible for "pulling down" the energy of the native structure in these models.

However, the basic interactions in proteins are the same as in a random heteropolymer; therefore it was suggested in [18] that the distinguishing feature of biologi-

cally active molecules was some kind of *sequence* optimization to allow folding to a unique structure, i.e., to achieve the energy gap. Further, a Metropolis Monte Carlo algorithm was introduced in sequence space to design sequences which possess such an energy gap [18,19]. The sequence design problem is often called the "inverse folding problem" in the following sense. In the folding problem the sequence is known and is quenched but chain conformation undergoes fluctuations until it reaches the native state (global energy minimum). In the design problem the structure (target conformation) is known and quenched, but the *sequence* is allowed to change in order that the target conformation have a low energy and that it be separated by an energy gap from other conformations.

The annealing in sequence space was done in [18,19] by Metropolis Monte Carlo procedure. A "selective temperature" $T_{\rm sel}$ was set in sequence space and the random heteropolymer was mutated at this "temperature" with the constraint that the amino acid composition remain unaltered to prevent the sequence from going to a homopolymer. The energy of a particular structure for a given sequence was defined in terms of the contact energies of the monomers. It was asserted in [18] that this design procedure generates a canonical distribution in sequence space with a temperature $T_{\rm sel}$ so that the probability of occurrence of any sequence, in this algorithm, is just the Boltzmann weight depending on the energy of that sequence in the native state and the selective temperature. The physics of the system in sequence space is identical to a ferromagnetic Ising system on an inhomogeneous lattice (corresponding to the target conformation) at the selective temperature with the constraint that the magnetization be constant. This equivalence implies that phase transitions are possible in sequence space which correspond to the dominance of very nonrandom sequences.

However, the constraint of constant composition does not matter in the thermodynamic limit since a phase transition in sequence space leads to thermodynamically large "hydrophobic" rich and "hydrophilic" rich regions—just like spin up and spin down domains in the Ising model. The condition of constant composition gives rise to an interface between hydrophobic and hydrophilic regions the energy of which is nonextensive and is not essential in thermodynamic limit.

Since the design procedure generates an ensemble of sequences which are certainly nonrandom, it is natural to expect that statistical mechanics of such sequences will be unusual. Taking into account the importance of designed sequences as a model for proteins it makes it interesting from both the physical and the biophysical points of view to study such heteropolymers.

The subject of the present paper is an analytical study of statistical-mechanical properties of heteropolymers with designed sequences. The calculations are sketched in Sec. II and the results are discussed in Sec. III.

II. THE MODEL AND CALCULATIONS

The interaction term in the Hamiltonian can be taken in a standard form:

$$\mathcal{H} = \frac{1}{2}\sum_{i,j}^{N} b_{ij} U(\mathbf{r_i} - \mathbf{r_j}), \tag{2.1}$$

where the conformation of the polymer is described by the coordinates of its monomers $\{\mathbf{r_i}\}$ and $U(\mathbf{r_i} - \mathbf{r_j})$ is a short-range potential. N is the number of monomers in the chain. The binary interaction virial coefficients are given by [20]

$$b_{ij} = 2\big[b_0 + A(\sigma_i + \sigma_j) + \chi\sigma_i\sigma_j\big]. \tag{2.2}$$

The sequence of monomers is described by the variables $\{\sigma_i\}$. $\sigma_i = 1$ if monomer i is of type A (say, hydrophobic) and $\sigma_i = -1$ if it is of type B (hydrophilic). When the interactions between similar monomers are equal ($b_{AA} = b_{BB}$) then $A = 0$. The composite Flory parameter $\chi = (b_{AA} + b_{BB})/2 - b_{AB}$ will be negative in the case of interest where similar monomers attract each other. $b_0 < 0$ provides an average attraction between monomers; this sequence-nonspecific term shifts equilibrium towards globular states. Thus for the calculation of thermodynamic properties one needs to consider only compact globular states.

The target (native) structure is defined via its coordinates $\{\mathbf{r}_i^0\}$. Thus the energy of the sequence $\{\sigma_i\}$ folded to this structure is given by

$$\mathcal{H}_0(\{\sigma_i\}) = \frac{1}{2}\sum_{i,j}^{N} b_{ij} U(\mathbf{r}_i^0 - \mathbf{r}_j^0). \tag{2.3}$$

As mentioned earlier, the selection procedure, which is the Monte Carlo optimization in sequence space, is nothing but the selection of the sequence depending on its energy in the target conformation and the "temperature" in sequence space. This converges to the canonical ensemble of sequences in sequence space with their energies given by that in the target conformation. Thus in this model the probability of occurrence of a particular sequence $\{\sigma_i\}$ is given by [18]

$$P\{\sigma_i\} = \frac{1}{\tilde{Z}} \exp\left(-\frac{\mathcal{H}_0(\{\sigma_i\})}{T_{\rm sel}}\right), \tag{2.4}$$

where

$$\tilde{Z} = \sum_{\{\sigma_i\}} \exp\left(-\frac{\mathcal{H}_0(\{\sigma_i\})}{T_{\rm sel}}\right) \tag{2.5}$$

and $T_{\rm sel}$ is the selective temperature in sequence space. To calculate the free energy within the framework of this model one needs to average the free energy over all possible sequences with the probability distribution given by Eq. (2.4), i.e.,

$$\langle F\rangle_{\rm av} = -kT \sum_{\{\sigma_i\}} \ln Z\{\sigma_i\} P\{\sigma_i\}, \tag{2.6}$$

where $\langle\cdots\rangle_{\rm av}$ denotes averaging over all possible sequences $\{\sigma_i\}$ with a probability distribution for their occurrence $P\{\sigma_i\}$ given by Eq. (2.4). The appearance of $P\{\sigma_i\}$ in the averaging procedure in Eq. (2.6) is the main

feature which distinguishes selected sequences from random ones. In the latter case each sequence is taken with *a priori* probability 2^{-N} while for selected sequences the weight factor P biases the averaging to take into account sequences which fit the native state with low energies.

Z in Eq. (2.6) is the configurational partition function of a copolymer with a given sequence:

$$Z = \sum_{\{r_i\}} \left[\exp\left(-\frac{\mathcal{H}(\{r_i\})}{k_b T} \right) \right] \prod_i g(\mathbf{r}_{i+1} - \mathbf{r}_i). \quad (2.7)$$

The summation is over all conformations of a protein which are expressed through the coordinates of the monomers. The g functions describe the covalent structure of the chain. They impose restrictions on the mutual positions of monomers which are nearest neighbors along the chain. The standard Gaussian form was suggested for these functions in [21]:

$$g(\mathbf{r}_{j+1}^{\alpha} - \mathbf{r}_j^{\alpha}) = \frac{1}{(2\pi a^2)^{3/2}} \exp\left[-\frac{(\mathbf{r}_{j+1}^{\alpha} - \mathbf{r}_j^{\alpha})^2}{2a^2} \right]. \quad (2.8)$$

In order to average the free energy over all possible sequences, as suggested by Eq. (2.6), one resorts to the replica method. This requires averaging the nth power of the partition function which is given by

$$\langle Z^n \rangle_{\text{av}} = \sum_{\{\sigma_i\}} \int \mathcal{D}\mathbf{r}_j^{\alpha} g(\mathbf{r}_{j+1}^{\alpha} - \mathbf{r}_j^{\alpha}) e^{-\frac{b_0}{T} \sum_{ij} U_{ij}^{\alpha}} \times \exp\left[-\sum_{\alpha} \sum_{i,j} \frac{\chi}{T} \sigma_i U(\mathbf{r}_i^{\alpha} - \mathbf{r}_j^{\alpha}) \sigma_j \right] P\{\sigma_i\}. \quad (2.9)$$

$U_{ij}^{\alpha} = U(\mathbf{r}_i^{\alpha} - \mathbf{r}_j^{\alpha})$ with $\mathbf{r}_i^{\alpha}$ the position of the ith monomer in replica α. We can rewrite Eq. (2.9) as

$$\left\langle \int \mathcal{D}\mathbf{r}_j^{\alpha} g(\mathbf{r}_{j+1}^{\alpha} - \mathbf{r}_j^{\alpha}) e^{-\frac{b_0}{T} \sum_{ij} U_{ij}^{\alpha}} \exp\left[b \sum_{\alpha} \int d\mathbf{R}_1 d\mathbf{R}_2 \sum_i \sigma_i \delta(\mathbf{r}_i^{\alpha} - \mathbf{R}_1) U(\mathbf{R}_1 - \mathbf{R}_2) \sum_j \sigma_j \delta(\mathbf{r}_j^{\alpha} - \mathbf{R}_2) \right] \right\rangle_{\text{av}}, \quad (2.10)$$

with $b = -\chi/T$ to be positive in the case of interest, although this approach can be generalized to either sign of b. We want to study the statistical properties of the system independent of the particular target structure selected. Thus we need to average over all possible target structures and all possible sequences. Performing the average over the variables $\{\sigma_i\}$ and $\{\mathbf{r}_j^0\}$ such that only compact target conformations are averaged over one obtains

$$\left\langle \int \mathcal{D}\mathbf{r}_j^{\alpha} \mathcal{D}\mathbf{r}_j^0 g(\mathbf{r}_{j+1}^{\alpha} - \mathbf{r}_j^{\alpha}) g(\mathbf{r}_{j+1}^0 - \mathbf{r}_j^0) e^{-\frac{b_0}{T} \sum_{ij} U_{ij}^{\alpha}} e^{-\frac{b_0}{T_{\text{sel}}} \sum_{ij} U_{ij}^0} \times \exp\left[b \sum_{\alpha} \int d\mathbf{R}_1 d\mathbf{R}_2 \sum_i \sigma_i \delta(\mathbf{r}_i^{\alpha} - \mathbf{R}_1) U(\mathbf{R}_1 - \mathbf{R}_2) \sum_j \sigma_j \delta(\mathbf{r}_j^{\alpha} - \mathbf{R}_2) + b_s \int d\mathbf{R}_1 d\mathbf{R}_2 \sum_i \sigma_i \delta(\mathbf{r}_i^0 - \mathbf{R}_1) U(\mathbf{R}_1 - \mathbf{R}_2) \sum_j \sigma_j \delta(\mathbf{r}_j^0 - \mathbf{R}_2) \right] \right\rangle_{\text{av}}, \quad (2.11)$$

with $b_s = -\chi/T_{\text{sel}}$, T_{sel} being the selective temperature. The second exponent in the above expression can be written as

$$\exp\left[\sum_{\alpha} b_{\alpha} \int d\mathbf{R}_1 d\mathbf{R}_2 \sum_i \sigma_i \delta(\mathbf{r}_i^{\alpha} - \mathbf{R}_1) U(\mathbf{R}_1 - \mathbf{R}_2) \sum_j \sigma_j \delta(\mathbf{r}_j^{\alpha} - \mathbf{R}_2) \right] \quad (2.12)$$

by defining $b_{\alpha} = b$ for $\alpha = (1, ..., n)$ and $b_{\alpha} = b_s$ for $\alpha = 0$.

By performing a Hubbard-Stratonovich transformation of the variables

$$\sum_i \sigma_i \delta(\mathbf{r}_i^{\alpha} - \mathbf{R}) \ \forall \alpha \in (0, \ldots, n)$$

the second exponential of Eq. (2.11) can be rewritten as

$$\int \mathcal{D}\Psi_{\alpha}(\mathbf{R}) \exp\left[-\sum_{\alpha} \frac{1}{4b_{\alpha}} \int d\mathbf{R}_1 d\mathbf{R}_2 \Psi_{\alpha}(\mathbf{R}_1) \Psi_{\alpha}(\mathbf{R}_2) U^{-1}(\mathbf{R}_1 - \mathbf{R}_2) + \sum_{\alpha} \int d\mathbf{R} \Psi_{\alpha}(\mathbf{R}) \sum_i \sigma_i \delta(\mathbf{r}_i^{\alpha} - \mathbf{R}) \right]. \quad (2.13)$$

Now performing the trace over σ_i one obtains

$$\langle Z^n \rangle_{\text{av}} = \frac{1}{\bar{Z}} \left\langle \int \mathcal{D}\Psi_{\alpha} \mathcal{D}\Psi_0 \exp\left[-\sum_{\alpha=0}^{n} \frac{1}{4b_{\alpha}} \int d\mathbf{R}_1 d\mathbf{R}_2 \Psi_{\alpha}(\mathbf{R}_1) \Psi_{\alpha}(\mathbf{R}_2) U^{-1}(\mathbf{R}_1 - \mathbf{R}_2) + \sum_i \ln \cosh\left(\sum_{\alpha} \int d\mathbf{R} \Psi_{\alpha}(\mathbf{R}) \delta(\mathbf{r}_i^{\alpha} - \mathbf{R}) \right) \right] \right\rangle_{\text{th}}, \quad (2.14)$$

where $\langle\rangle_{\rm th}$ indicates the integral over the $\{\mathbf{r}_i^0\}$ and the $\{\mathbf{r}_i^\alpha\}$ variables including the g factors. Expanding lncosh to fourth order in $\Psi_\alpha(R)$ the last exponential in the above equation can be rewritten as

$$\exp\Big[-\sum_{\alpha=0}^{n}\frac{1}{4b_\alpha}\int d\mathbf{R}_1 d\mathbf{R}_2\Psi_\alpha(\mathbf{R}_1)\Psi_\alpha(\mathbf{R}_2)U^{-1}(\mathbf{R}_1-\mathbf{R}_2)$$
$$+\frac{1}{2}\sum_{\alpha\beta=0}^{n}\int d\mathbf{R}_1 d\mathbf{R}_2\Psi_\alpha(\mathbf{R}_1)\Psi_\beta(\mathbf{R}_2)\Big(\sum_i\delta(\mathbf{r}_i^\alpha-\mathbf{R}_1)\delta(\mathbf{r}_i^\beta-\mathbf{R}_2)\Big)$$
$$-\frac{1}{12}\sum_{\alpha\beta\gamma\delta=0}^{n}\int d\mathbf{R}_1 d\mathbf{R}_2 d\mathbf{R}_3 d\mathbf{R}_4\Psi_\alpha(\mathbf{R}_1)\Psi_\beta(\mathbf{R}_2)\Psi_\gamma(\mathbf{R}_3)\Psi_\delta(\mathbf{R}_4)$$
$$\times\Big(\sum_i\delta(\mathbf{r}_i^\alpha-\mathbf{R}_1)\delta(\mathbf{r}_i^\beta-\mathbf{R}_2)\delta(\mathbf{r}_i^\gamma-\mathbf{R}_3)\delta(\mathbf{r}_i^\delta-\mathbf{R}_4)\Big)\Big], \tag{2.15}$$

setting

$$\Phi_\alpha(\mathbf{R}_2)=\frac{1}{2b_\alpha}\int d\mathbf{R}_1 U^{-1}(\mathbf{R}_1-\mathbf{R}_2)\Psi_\alpha(\mathbf{R}_1) \tag{2.16}$$

one gets for the n replica partition function averaged over the sequence variables $\{\sigma\}$ as

$$\langle Z^n\rangle_{\rm av}=\frac{1}{\bar{Z}}\Big\langle\int\mathcal{D}\Phi_\alpha\mathcal{D}\Phi_0\exp\Big[-\sum_{\alpha=0}^{n}b_\alpha\int d\mathbf{R}_1 d\mathbf{R}_2\Phi_\alpha(\mathbf{R}_1)\Phi_\alpha(\mathbf{R}_2)U(\mathbf{R}_1-\mathbf{R}_2)$$
$$+2\sum_{\alpha\beta=0}^{n}b_\alpha b_\beta\int d\mathbf{R}_1 d\grave{\mathbf{R}}_1\Phi_\alpha(\grave{\mathbf{R}}_1)U(\mathbf{R}_1-\grave{\mathbf{R}}_1)\int d\mathbf{R}_2 d\grave{\mathbf{R}}_2\Phi_\alpha(\grave{\mathbf{R}}_2)U(\mathbf{R}_2-\grave{\mathbf{R}}_2)\mathbf{Q}_{\alpha\beta}(\mathbf{R}_1-\mathbf{R}_2)$$
$$-\frac{4}{3}\sum_{\alpha\beta\gamma\delta}b_\alpha b_\beta b_\gamma b_\delta\int d\mathbf{R}_1 d\grave{\mathbf{R}}_1\cdots d\mathbf{R}_4 d\grave{\mathbf{R}}_4\Phi_\alpha(\grave{\mathbf{R}}_1)U(\mathbf{R}_1-\grave{\mathbf{R}}_1)\cdots\Phi_\delta(\grave{\mathbf{R}}_4)U(\mathbf{R}_4-\grave{\mathbf{R}}_4)$$
$$\times\sum_i\delta(\mathbf{r}_i^\alpha-\mathbf{R}_1)\delta(\mathbf{r}_i^\beta-\mathbf{R}_2)\delta(\mathbf{r}_i^\gamma-\mathbf{R}_3)\delta(\mathbf{r}_i^\delta-\mathbf{R}_4)\Big]\Big\rangle_{\rm th}, \tag{2.17}$$

where once again the integrals over $\{\mathbf{r}_i^0\}$ and $\{\mathbf{r}_i^\alpha\}$ are not explicitly written and $\alpha=0$ denotes the target conformation and $\alpha=(1,\ldots,n)$ denote the n replicas.

$$\mathbf{Q}_{\alpha\beta}(\mathbf{R}_1-\mathbf{R}_2)=\sum_i\delta(\mathbf{r}_i^\alpha-\mathbf{R}_1)\delta(\mathbf{r}_i^\beta-\mathbf{R}_2). \tag{2.18}$$

The parameter $Q_{\alpha\beta}$ is a measure of the overlap between conformations α and β $\forall\alpha,\beta\in(0,\ldots,n)$. If the two folds α and β are completely different then $Q_{\alpha,\beta}=0$. In the opposite extreme, if the two folds are identical then $Q_{\alpha\beta}=\rho\delta(\mathbf{R}_1-\mathbf{R}_2)$ where ρ is the density. In particular $Q_{0\alpha}$ represents the overlap of the replica α with the target conformation. We can evaluate Eq. (2.17) by switching to the $Q_{\alpha\beta}$ variables. The corresponding entropy for this change of variable is

$$\ln S\{Q_{\alpha\beta}\}=\Big\langle\delta(Q_{\alpha\beta}(\mathbf{R}_1-\mathbf{R}_2)-\sum_i\delta(\mathbf{r}_i^\alpha-\mathbf{R}_1)\delta(\mathbf{r}_i^\beta-\mathbf{R}_2))\Big\rangle_{\rm th}. \tag{2.19}$$

Let us consider the overlap between the conformation α and the native state in particular.

(1) $Q_{0,\alpha}=0$. This is when there is no overlap between the replicas and the target conformation. In this case one obtains following the arguments in [8] that there is a one step replica symmetry breaking in $Q_{\alpha\beta}$ $\forall\alpha,\beta\in(1,\ldots,n)$ and we can construct a Parisi-type hierarchical matrix for the order parameter with the form

$$Q_{\alpha\beta}=\begin{cases}\rho\delta(\mathbf{R}_1-\mathbf{R}_2) & \text{for } \alpha,\beta \text{ in the same group}\\ 0 & \text{for } \alpha,\beta \text{ in different groups.}\end{cases} \tag{2.20}$$

(2) the fold α and the target conformation are identical then $Q_{0\alpha}(\mathbf{R}_1-\mathbf{R}_2)=\rho\delta(\mathbf{R}_1-\mathbf{R}_2)$ where ρ is the density. In this case $Q_{\alpha\beta}(\mathbf{R}_1-\mathbf{R}_2)=\rho\delta(\mathbf{R}_1-\mathbf{R}_2)$ for all α and β and all the replica folds are identical to the target structure to within a microscopic scale.

(3) We can express the intermediate case of similarity between the fold α and the target state by writing the order parameter in terms of a function $\varphi_{0\alpha}$ with unit scale as

$$Q_{0\alpha}(\mathbf{R}_1-\mathbf{R}_2)=\frac{\rho}{R_t^{0\,3}}\varphi_{0\alpha}\left(\frac{\mathbf{R}_1-\mathbf{R}_2}{R_t^0}\right), \tag{2.21}$$

with $\int Q_{0\alpha}(\mathbf{R}_1-\mathbf{R}_2)d\mathbf{R}_1 d\mathbf{R}_2=N$. This means that replica α repeats the target fold within some scale of fluctuations R_t^0. The other order parameter in the problem is the overlap between different replicas. This is repre-

sented by $Q_{\alpha\beta}(\mathbf{R}_1 - \mathbf{R}_2), \forall \alpha\beta \in (1, \ldots, n)$ which has the characteristic length scale R_t. This length scale, which is the measure of the overlap between different replicas, represents the width of the tube that the replicas are confined to. Assuming a one step replica symmetry breaking in this order parameter, R_t is either zero or infinity. The free energy corresponds to the free energy of confinement of replicas of the same group in a tube of radius R_t and further confinement of these tubes in a tube of radius R_t^0. The corresponding entropy loss scales as $-1/R_t^{0\,2}$ [2,3]. The energy in terms of $Q_{\alpha\beta}(\mathbf{R})$ is found by integration of the Gaussian integral (2.17).

$$\left\langle \int \mathcal{D}\Phi_\alpha(\mathbf{k}) \exp\left[-V \sum_{\alpha,\beta=0}^{n} \sum_{\mathbf{k}} [b_\alpha \delta_{\alpha\beta} - 2 b_\alpha b_\beta Q_{\alpha\beta}(\mathbf{k})] \Phi_\alpha(\mathbf{k}) \Phi_\beta(-\mathbf{k}) \right] \right\rangle_{\text{th}}, \tag{2.22}$$

where V is the volume of the system, $\mathbf{k}$ is the wave vector, and $Q_{\alpha\beta}(\mathbf{k})$ and $\Phi(\mathbf{k})$ are the Fourier transforms of the order parameters. Performing the integral over the fields $\Phi_\alpha(\mathbf{k}), \alpha = 0, ..., n$ one obtains

$$\int d\mathbf{k} \ln[\det \mathbf{P}_{\alpha\beta}(\mathbf{k})], \tag{2.23}$$

with $P_{\alpha\beta}(\mathbf{k}) = b_\alpha \delta_{\alpha\beta} - 2 b_\alpha b_\beta Q_{\alpha\beta}(\mathbf{k})$. $P_{\alpha\beta}(\mathbf{k})$ is a symmetric matrix. We can rescale $\mathbf{k}$ to $R_t^0 \mathbf{k}$. Thus

$$\frac{1}{R_t^{0\,3}} \int d(R_t^0 \mathbf{k}) \ln[\det P_{\alpha\beta}(R_t^0 \mathbf{k})]. \tag{2.24}$$

Thus the free energy per monomer as a function of R_t^0 is expected to have the form [8]

$$\frac{\mathcal{F}(R_t^0)}{Nn} = -\frac{A_1}{R_t^{0\,2}} + \frac{A_2}{R_t^{0\,3}}, \tag{2.25}$$

with $A_1, A_2 > 0$.

Minimizing the free energy with respect to R_t^0 we find that R_t^0 must either be infinite or must be a, the microscopic length scale in the problem. Thus

$$Q_{0\alpha} = \begin{cases} \rho\delta(\mathbf{R}_1 - \mathbf{R}_2) & \text{for replica } \alpha \text{ identical to the target structure, i.e., } R_t^0 = 0 \\ 0 & \text{for } \alpha \text{ different from the target structure is } R_t^0 = \infty. \end{cases} \tag{2.26}$$

Thus we see that either case (1) or case (2) applies and that $Q_{0\alpha}$ does not vary continuously but can take on only two possible values. This corresponds to the sharp change of the state of the system from the disordered globule or frozen state to the target conformation depending on the actual temperature and the selective temperature. The temperatures at which these transitions take place can be found by just comparing the free energies of the various states in question. We do this for the case when the interaction potential between the monomers is a δ function representing contact interactions.

A. The free energy of the target state

The free energy of the target state at a given selective temperature can be calculated in the mean field using the saddle point approximation to perform the integrals over the Ψ_α fields. This is identical to the calculation of the free energy of a ferromagnetic system on an inhomogeneous lattice at a given temperature. As is well known, the system undergoes second-order phase transition at a particular temperature below which there is a nonzero value for the magnetization. Following Eq. (2.9) we get for the n replica partition function

$$\langle Z^n \rangle_{\text{av}} = \left\langle \int \mathcal{D}\mathbf{r}_j^0 g(\mathbf{r}_{j+1}^0 - \mathbf{r}_j^0) e^{-(\frac{b_0}{T} + \frac{b_0}{T_{\text{sel}}}) \sum_{ij} U_{ij}^0} \exp\left[\sum_{i,j} -(b_s + nb) \sigma_i U(\mathbf{r}_i^0 - \mathbf{r}_j^0) \sigma_j \right] \right\rangle_{\text{av}} \tag{2.27}$$

since in this case $\mathbf{r}_i^\alpha$ is identical to $\mathbf{r}_i^0$ for all α and i. One can once again perform the Hubbard-Stratonovich transformation as done in going from Eq. (2.11) to Eq. (2.15) except that in this case the fields Ψ_α are identical for all α. Thus the expression corresponding to Eq. (2.14) for the partition function is obtained by simply replacing all the $\Psi_\alpha = \Psi_0$,

$$\langle Z^n\rangle_{\rm av} = \frac{1}{\tilde{Z}}\left\langle \int \mathcal{D}\Psi_0 \exp\left[-\frac{1}{4(nb+b_s)}\int d\mathbf{R}_1 d\mathbf{R}_2 \Psi_0(\mathbf{R}_1)\Psi_0(\mathbf{R}_2)U^{-1}(\mathbf{R}_1-\mathbf{R}_2) + \int d\mathbf{R}\Psi_0(\mathbf{R})\sum_i \sigma_i \delta(\mathbf{r}_i^0-\mathbf{R})\right]\right\rangle_{\rm th}, \tag{2.28}$$

where once again the integrals over $\{\mathbf{r}_i^0\}$ and $\{\mathbf{r}_i^\alpha\}$ are not explicitly written. For $U(\mathbf{R}_1-\mathbf{R}_2)=\delta(\mathbf{R}_1-\mathbf{R}_2)$ the above equation becomes

$$\langle Z^n\rangle_{\rm av} = \frac{1}{\tilde{Z}}\left\langle \int \mathcal{D}\Psi_0 \exp\left[-\frac{1}{4(nb+b_s)}\int d\mathbf{R}\Psi_0^2(\mathbf{R}) + \frac{\rho}{2}\int d\mathbf{R}\Psi_0(\mathbf{R})\right]\right\rangle_{\rm th}. \tag{2.29}$$

From this expression the energy of the target state can be easily found in the mean-field approximation by performing the Gaussian integrals over the field Ψ_0 using the method of steepest descent and taking the limit $n \to 0$ as is required by the replica method [22]. Note that $\tilde{Z}$ in the above case is equal to $\ln(1-2b_s\rho)$ when $2b_s\rho < 1$, i.e., $T_s > -2\chi\rho$. It is found that for $T_s > -2\chi\rho$ for the field Ψ_0 the mean value of the field is zero while for $T_s < -2\chi\rho$ the field Ψ_0 has a nonzero mean value. In this case the fourth-order term has also to be taken into account. This is due to the phase transition in the sequence space of the monomers on the inhomogeneous target conformation. It is identical to the phase transitions seen in magnets at the critical point below which there is nonzero magnetization and the saddle point integration has to be performed about the new minimum that appears. In either case the energy density of the target conformation is calculated to be

$$\mathcal{F} = \begin{cases} -(2b\rho)/(1-2b_s\rho), & T_s > -2\chi\rho \\ -\frac{3}{16}(b)/(\rho b_s^2)\left[1-1/(2b_s\rho)\right], & T_s < -2\chi\rho. \end{cases} \tag{2.30}$$

To calculate the free energy density relative to the disordered globule we must calculate the entropy loss in constraining the replicas to be identical to the target conformation to within a microscopic scale. The replicas repeat the target conformation to within a scale of $R_t^0 v^{1/3}$. We know that because of polymeric effect, after placing one monomer we must place the next one in a volume a^3. There are a^3/v ways of doing this so the corresponding entropy is $\ln(a^3/v)$ per monomer. Thus the total loss entropy loss for the n replicas is

$$S = Nn\ln(v/a^3). \tag{2.31}$$

Thus the free energy density of the target state is given by

$$\mathcal{F} = \begin{cases} -(2b\rho)/(1-2b_s\rho) + \ln(v/a^3), & T_s > -2\chi\rho \\ -\frac{3}{16}(b)/(\rho b_s^2)\left[1-1/(2b_s\rho)\right] + \ln(v/a^3), & T_s < -2\chi\rho. \end{cases} \tag{2.32}$$

B. The free energy of the disordered globule and the frozen globular state

As was noted earlier, when there is no overlap with the native state the problem becomes identical to the problem of the one of the random heteropolymer studied in [8].

Setting $Q_{0\alpha}=0$ in Eq. (2.15) and including only up to the second-order terms in $\Psi_\alpha(R)$ one finds the n replica partition function to be

$$\langle Z^n\rangle_{\rm av} = \frac{1}{\tilde{Z}}\left\langle \int \mathcal{D}\Psi_\alpha \mathcal{D}\Psi_0 \exp\left[-\sum_{\alpha=0}^{n}\frac{1}{4b_\alpha}\int d\mathbf{R}_1 d\mathbf{R}_2 \Psi_\alpha(\mathbf{R}_1)\Psi_\alpha(\mathbf{R}_2)U^{-1}(\mathbf{R}_1-\mathbf{R}_2) + \frac{1}{2}\sum_{\alpha\beta=1}^{n}\int d\mathbf{R}_1 d\mathbf{R}_2 \Psi_\alpha(\mathbf{R}_1)\Psi_\beta(\mathbf{R}_2)Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2) + \frac{\rho}{2}\int d\mathbf{R}_1 \Psi_0^2(\mathbf{R}_1)\right]\right\rangle_{\rm th}. \tag{2.33}$$

The integral over the fields $\Psi_0(R)$ cancels with the terms in $\tilde{Z}$ and one finds that there is no coupling between the $\Psi_\alpha(R)$ and $\Psi_0(R)$ fields when the fluctuations in the order parameter $Q_{0\alpha}$ are neglected. Following the calculations in [8], assuming a one step replica symmetry in $Q_{\alpha\beta}$ with x_0 as the variational parameter, one obtains for the free energy

$$\frac{f}{n} = \frac{\ln[1-2b\rho x_0]}{x_0} - \frac{s}{x_0}, \tag{2.34}$$

where $s = \ln(a^3/v) > 0$ denotes the flexibility of the chain. Maximizing with respect to x_0 and solving for x_0 to lowest order in s one obtains

$$x_0 = \begin{cases} \frac{T\sqrt{s}}{2(-\chi)\rho} & \text{for } T < T^c = \frac{2(-\chi)\rho}{\sqrt{s}} \\ 1 & \text{for } T > T^c. \end{cases} \tag{2.35}$$

The fluctuations of the $Q_{0\alpha}$ order parameter will affect the freezing temperature. In order to investigate the effect of fluctuations we must introduce a finite range

into the potential and set $U(\mathbf{k}) = 1 - c^2\mathbf{k}^2$ where c^2 is a surface tension coefficient which suppresses large wave vectors $\mathbf{k}$. We then set $b_\alpha(\mathbf{k}) = b_\alpha(1-c^2\mathbf{k}^2)$. We consider the partition function in the form of Eq. (2.17).

We group the replicas according to the replica symmetry breaking (RSB) pattern [8]. For replicas in the same group we replace the quantity $\sum_i \delta(\mathbf{r}_i^\alpha - \mathbf{R}_1)\delta(\mathbf{r}_i^\beta - \mathbf{R}_2)$ by $\rho\delta(\mathbf{R}_1 - \mathbf{R}_2)$. On the mean-field level we will have $\langle\sum_i \delta(\mathbf{r}_i^\alpha - \mathbf{R}_1)\delta(\mathbf{r}_i^0 - \mathbf{R}_2)\rangle_{\rm th} = 0$ for the replicas and the target conformation. In order to investigate fluctuations of the $Q_{\alpha 0}$ order parameter and interactions between replicas in different groups and the target conformation we need to expand the bilinear term $\Phi_\alpha(\mathbf{k})\Phi_0(-\mathbf{k})$ for replica α and the target conformation as follows:

$$\frac{1}{\tilde{Z}}\left\langle \int \mathcal{D}\Phi_\alpha(\mathbf{R})\mathcal{D}\Phi_0(\mathbf{R}) \exp\Big[2bb_s \sum_{(A)}\sum_i \sum_{\alpha\in A}\int d\mathbf{R}_1 d\mathbf{R}_1' \Phi_\alpha(\mathbf{R}_1')U(\mathbf{R}_1 - \mathbf{R}_1')\delta(\mathbf{r}_i^\alpha - \mathbf{R}_1) \times \int d\mathbf{R}_2 d\mathbf{R}_2' \Phi_0(\mathbf{R}_2')U(\mathbf{R}_2 - \mathbf{R}_2')\delta(\mathbf{r}_i^\beta - \mathbf{R}_2)\Big]\right\rangle_{\rm th}, \tag{2.36}$$

with (A) indicating a sum over all the possible groups of replicas. By Fourier transformation

$$\frac{1}{\tilde{Z}}\left\langle \int \mathcal{D}\Phi_\alpha(\mathbf{k}) \exp\Big[-V\sum_{\alpha=1}^{n}\sum_{\mathbf{k}\neq 0} b(\mathbf{k})\Phi_\alpha(\mathbf{k})\Phi_\alpha(-\mathbf{k}) - V\sum_{\mathbf{k}\neq 0} b_s(\mathbf{k})\Phi_0(\mathbf{k})\Phi_0(-\mathbf{k}) + 2\sum_\alpha\sum_i\sum_{\mathbf{k}_1,\mathbf{k}_2\neq 0} b(\mathbf{k}_1)e^{i\mathbf{k}_1\mathbf{r}_i^\alpha}\Phi_\alpha(\mathbf{k}_1)b_s(\mathbf{k}_2)e^{i\mathbf{k}_2\mathbf{r}_i^0}\Phi_0(\mathbf{k}_2)\Big]\right\rangle_{\rm th}. \tag{2.37}$$

We expand the bilinear term in the replica fields and the target field. The second-order term vanishes after the thermal average. At fourth order, only terms with a pair of replicas from the same group and a pair in the target state will survive. Hence from Eq. (2.17) we obtain

$$\frac{1}{\tilde{Z}}\left\langle \int \mathcal{D}\Phi_\alpha(\mathbf{k}) \exp\Big[-V\sum_A\sum_{\alpha\in A}\sum_{\mathbf{k}\neq 0} P_{\alpha 0}(\mathbf{k})\Phi_\alpha(\mathbf{k})\Phi_0(-\mathbf{k})\Big] \times\Big\{1 + \sum_{(A)}\sum_{\alpha\gamma\in A}\sum_{i,j}\sum_{\mathbf{k}_1\ldots\mathbf{k}_4\neq 0} b(\mathbf{k}_1)b_s(\mathbf{k}_2)b(\mathbf{k}_3)b_s(\mathbf{k}_4) \times e^{i(\mathbf{k}_1\mathbf{r}_i^\alpha + \mathbf{k}_3\mathbf{r}_j^\gamma)}\Phi_\alpha(\mathbf{k}_1)\Phi_\gamma(\mathbf{k}_3)e^{i(\mathbf{k}_2\mathbf{r}_i^0 + \mathbf{k}_4\mathbf{r}_j^0)}\Phi_0(\mathbf{k}_2)\Phi_0(\mathbf{k}_4)\Big\}\right\rangle_{\rm th}, \tag{2.38}$$

where $P_{\alpha 0}(\mathbf{k}) = -2bb_s(\mathbf{k})\rho$, $P_{\alpha\alpha}(\mathbf{k}) = b(k) - 2b^2(\mathbf{k})\rho$, and $P_{00}(\mathbf{k}) = b_s(k) - 2b_s^2(\mathbf{k})\rho$. The first integral in Eq. (2.34) is the Gaussian integral denoted as $C(x_0)$. The integral of the fourth-order term is calculated to be

$$C(x_0)\left\langle 4\sum_{\mathbf{k}_1,\mathbf{k}_2}\sum_{(A)}\sum_{\alpha\gamma\in A}\sum_{i,j} b^2(\mathbf{k}_1)b_s^2(\mathbf{k}_2)e^{i\mathbf{k}_1(\mathbf{r}_i^\alpha - \mathbf{r}_j^\gamma)}e^{i\mathbf{k}_2(\mathbf{r}_i^\beta - \mathbf{r}_j^\delta)}[P^{-1}]^A_{\alpha\gamma}(\mathbf{k}_1)\frac{1}{b_s - 2b_s^2\rho}\right\rangle_{\rm th}. \tag{2.39}$$

We will represent the elements of the inverse Parisi matrix, $P^{-1}(\mathbf{k})$ as $p(\mathbf{k})$ for the off-diagonal elements and $\tilde{p}(\mathbf{k})$ for the diagonal elements where

$$p(\mathbf{k}) = \frac{\gamma(\mathbf{k})}{b(\mathbf{k})[1 - \gamma(\mathbf{k})x_0]} \quad \text{and} \quad \tilde{p}(\mathbf{k}) = \frac{1 + \gamma(\mathbf{k})(1 - x_0)}{b(\mathbf{k})[1 - \gamma(\mathbf{k})x_0]}, \tag{2.40}$$

with $\gamma(\mathbf{k}) = 2b(\mathbf{k})\rho$. Then Eq. (2.40) becomes

$$4C(x_0)\frac{n}{x_0}\sum_{\mathbf{k}_1,\mathbf{k}_2\neq 0} b^2(\mathbf{k}_1)b_s^2(\mathbf{k}_2)\Big[(x_0 - 1)x_0 p(\mathbf{k}_1) + x_0\tilde{p}(\mathbf{k}_1)\Big]\frac{1}{b_s - 2b_s^2\rho}\left\langle \sum_{i,j} e^{i\mathbf{k}_1(\mathbf{r}_i^\alpha - \mathbf{r}_j^\alpha)}e^{i\mathbf{k}_2(\mathbf{r}_i^0 - \mathbf{r}_j^0)}\right\rangle_{\rm th}. \tag{2.41}$$

By substitution of (2.35) in (2.36)

$$4nC(x_0)\sum_{\mathbf{k}_1,\mathbf{k}_2}\frac{b(\mathbf{k}_1)b_s(\mathbf{k}_2)}{[1 - \gamma(\mathbf{k}_1)x_0][1 - \gamma_s(\mathbf{k}_2)]} \times\left\langle \sum_{i,j} e^{i\mathbf{k}_1(\mathbf{r}_i^\alpha - \mathbf{r}_j^\alpha)}e^{i\mathbf{k}_2(\mathbf{r}_i^0 - \mathbf{r}_j^0)}\right\rangle_{\rm th}, \tag{2.42}$$

where $\gamma_s(\mathbf{k}) = 2b_s(\mathbf{k})\rho$. The coefficient of the Gaussian term changes sign for $T < T^a = -2\chi\rho$. Far away from T^a, which is the region we will be interested in, we can consider $\gamma(\mathbf{k})$ independent of $\mathbf{k}$. Thus

$$= \frac{4bb_sC(x_0)n}{(1 - \gamma x_0)(1 - \gamma_s)}\left\langle \sum_{i,j}\delta(\mathbf{r}_i^\alpha - \mathbf{r}_j^\alpha)\delta(\mathbf{r}_i^0 - \mathbf{r}_j^0)\right\rangle_{\rm th}. \tag{2.43}$$

The thermal (configurational) average in (2.43) is the number of common contacts different folds α and the target configuration have. The overlap is mainly due to the contacts of neighboring monomers which is neglected in the mean-field theory for $Q_{\alpha 0}$ and becomes important when the flexibility of the chain increases. We denote this overlap as $N\epsilon^0$ where ϵ^0 is the small parameter of our perturbation expansion. With this correction the expression for the free energy becomes

$$\frac{f(x_0)}{n} = \ln b + \frac{\ln(1-\gamma x_0)}{x_0} - \frac{s}{x_0} + \frac{\epsilon^0 \gamma \gamma_s}{(1-\gamma x_0)(1-\gamma_s)}. \tag{2.44}$$

Expanding to lowest nonvanishing order for x_0 and imposing $\partial f(x_0)/\partial x_0 = 0$ we obtain

$$x_0 = \frac{1}{\gamma}\left(\frac{s}{1-\epsilon^0\gamma_s/(1-\gamma_s)}\right)^{1/2} \tag{2.45}$$

and therefore the freezing temperature is

$$T^c = (-2\chi\rho/\sqrt{s})[1-\epsilon^0\gamma_s/(1-\gamma_s)]^{1/2} \tag{2.46}$$

$$= T^{c0}[1-\epsilon^0\gamma_s/(1-\gamma_s)]^{1/2}. \tag{2.47}$$

$x_0 = 1$ for $T > T^c$. Thus above T^c the polymer is in the continuous part of the spectrum. Below this temperature the chain goes to the frozen globular phase. The fluctuations $\langle Q_{\alpha 0}^2 \rangle$ due to overlaps between replicas and the target structure result in a decrease of the freezing temperature. The more flexible the polymer chain the more important the overlaps due to neighboring monomer contacts. Therefore with increasing flexibility the freezing temperature will decrease. The free energy of the disordered globule is given by Eq. (2.43) when $x_0 = 1$. This occurs when $T > T_c$ when the heteropolymer is in the continuous part of the energy spectrum. In this case though it is compact it does not have a unique structure and hence is termed as a disordered globule. Below the temperature T^a free energy density of the disordered globule can be obtained to be

$$f = -\frac{3}{16\rho^2}\left(\frac{1}{b} - 2\rho\right)^2. \tag{2.48}$$

To find the curve on the T vs $T_{\rm sel}$ phase diagram along which the transition occurs from the frozen state to the target state one equates the free energies given by Eqs. (2.31), (2.43), and (2.45). For small s, which is the case we are dealing with, we find that the transition occurs when $T_{\rm sel} > -2\chi\rho$. Thus we compare the free energies of the target state above the ferromagnetic transition temperature with that of the frozen state. This gives the curve in the T vs T_s along which the transition from the frozen state to the target conformation occurs. Consistently to first order in ϵ^0 this curve has the equation

$$T_s = -\frac{4\chi\rho}{3\sqrt{s}}\left(1 - \frac{3}{4}\epsilon^0\sqrt{s}\right). \tag{2.49}$$

We label this temperature as T_s^c. It is important to note that since entropy of both target state and frozen state are vanishing the transition is independent of temperature. Correspondingly, there is no energy jump upon this transition, and thermodynamically it is the second-order phase transition.

Similarly, to find the curve on the phase diagram along which the transition occurs from the disordered globule to the target state, one equates the free energies in Eqs. (2.25) and (2.29) to obtain the coexistence curve between the target state and disordered globule for $T_s < T_s^c$. Thus for $T_s < -2\chi\rho$

$$\frac{3}{16}\frac{b}{\rho b_s^2}\left(1 - \frac{1}{2b_s\rho}\right) + \ln(v/a^3) = -\frac{3}{16\rho^2}\left(\frac{1}{b} - 2\rho\right)^2. \tag{2.50}$$

This can be solved for very low temperatures and one obtains

$$T = \frac{4}{\rho}T_s^2. \tag{2.51}$$

At temperatures $-2\chi\rho < T_s < T_s^c - \epsilon$, i.e., T_s close to T_s^c, one similarly obtains to first order

$$T = -\left[\frac{9}{2(-\chi)}\left(1 - \frac{3\sqrt{s}}{2(-\chi)}\right)\right](T_s - T_s^c) + \frac{-2\chi\rho}{\sqrt{s}}\left(1 - \frac{3\epsilon\sqrt{s}}{4}\right) \tag{2.52}$$

When $T_{\rm sel}$ is close to $-2\chi\rho$ one expects the fluctuations in the magnetization to be important and expects a discontinuity in the slope of the curve along which transition from the disordered globule to the target conformation occurs at $-2\chi\rho$. In contrast to the transition between the target state and the frozen state the transition between the disordered globule and the target state involves an entropy loss of $\ln(a^3/v)$ per monomer [see Eq. (2.31)] connected with ordering of the chain in the target conformation. Correspondingly, this transition is thermodynamically a *first-order* phase transition.

III. RESULTS AND DISCUSSION

The results of our calculations are summarized in the phase diagram Fig. 1 For $T_{\rm sel} > T_s^c$ one finds that on reducing the real temperature the disordered globule goes into the frozen state. At these selective temperatures the free energy of the target state is above the lower energy levels of the heteropolymer as a result of which the system at low temperatures goes into the frozen state [Fig 2(a)]. As the selective temperature is decreased the energy of the native state decreases and at $T_s = T_s^c$ it is exactly equal to the lowest frozen state [Fig 2(b)]. The fact that the selective temperature at which transition from the frozen state to the target state occurs is independent of the real temperature is because there is no entropy loss in this transition. Below T_s^c as one decreases the real temperature a transition occurs directly from the disordered globule state to the target state. At these temperatures the energy of the native state is much below the frozen

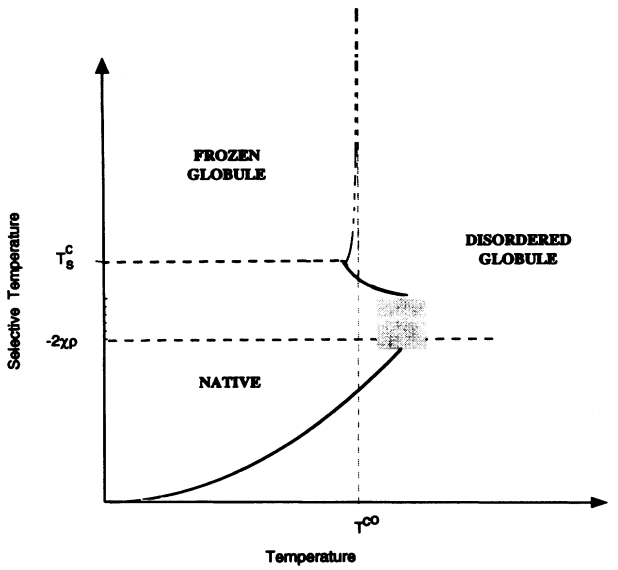

FIG. 1. The phase diagram shows the curves in the T_{sel}, T space along which transitions from one phase to another take place. For sequences generated at $T_{sel} > T_{sel}^c$ the polymer goes into the frozen state on cooling in real space while for sequences generated at $T_{sel} < T_{sel}^c$ the disordered globule goes into the target conformation on cooling. The fat line corresponds to first-order transition. The shaded area denotes the region near the "ferromagnetic transition point" in sequence space where fluctuations are strong, and mean-field theory described here is inapplicable. However, for sequences generated at lower selective temperature the transition temperature to the target conformation decreases due to increased correlations in sequences.

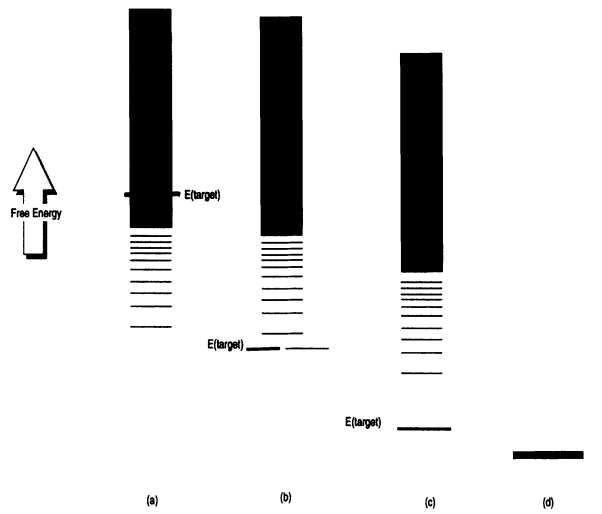

FIG. 2. (a) $T_{sel} > T_{sel}^c$, thus the energy of the target conformation is greater than that on the frozen states, hence on cooling the system goes into the frozen state. (b) $T_{sel} = T_{sel}^c$, this is transition selective temperature from the frozen state to the target state. (c) $T_{sel} < T_{sel}^c$ now, on cooling the system the system freezes into the target conformation (d) $T_{sel} = 0$. The statistical properties are dominated by chains with thermodynamically large homopolymeric segments and thus have a highly degenerate ground state and the system does not freeze into the target conformation.

states [Fig 2(c)]. Near $T_s^{c1} = -2\chi\rho$ the size of domains of monomer type A and that of monomer type B increases. Thus one gets large regions of A and B in the inhomogeneous "target" lattice, that is, the sequence average is dominated by sequences which when folded along the target conformation show A rich and B rich regions on this target lattice. At a given selective temperature T_s the correlations in composition as a function of the distance on the target lattice, that is, the chances of finding a monomer of the same type at a distance r are given by

$$C(r) = \exp[-r/\xi], \tag{3.1}$$

with $\xi = \alpha[T_s - (-2\chi\rho)]^{-\nu}$ where $\nu > 0$ is a critical exponent which need not be equal to that of the regular three-dimensional Ising model because of the inhomogeneity of the target lattice.

This results in the probability of a selected sequence having a contiguous region of length L of hydrophilic or hydrophobic monomers, i.e., homopolymerlike regions increase because of the increase in the correlation lengths of the sequence variables. The correlation length along the polymer will be 2ξ. As $T_s \to -2\chi\rho$ this correlation length diverges. As a result the transition temperature to the target conformation decreases.

For very low selective temperatures, the sequences that have a low energy in the target state have pronounced hydrophilic and hydrophobic contiguous regions. If the composition of the sequence was kept fixed one would obtain thermodynamically large domains of hydrophobic and hydrophilic regions and in the limit of zero selective temperature one would see total phase separation.

As a result the average property is dominated by sequences having large homopolymeric segments. Thus the temperature of transition to the native state further decreases. In the limit of zero selective temperature the average properties of the system are completely dominated by the homopolymeric sequence. Since the ground state of the homopolymer is infinitely degenerate, i.e., all conformations have the same energy, one finds that there is no transition to the target state even at zero temperature.

An analytical study of another model with selected native structure was published recently in [23,24]. Our analysis is consistent with this study as it reveals the same major phases—disordered, frozen, and the target state. However, quantitative comparison of our results with that of Sasai and Wolynes is difficult because the model used in [23,24] is different from ours. Indeed in the model of Sasai and Wolynes the target structure is singled out by "ultraspecific" forces which biased the chain towards the native structure. In our model *sequence selection* is the factor which distinguishes proteins from random heteropolymers, while basic interactions are assumed to be the same in proteins and any random copolymer made of monomers of the same type.

The results of the present analysis have several implications for protein folding. To make proper comparison with experiment we note that a simplified model is used where monomers are presented as beads and the side-chains with their degrees of freedom are ignored. Side-chain packing is very important in the native state

of proteins while it is lost in the so-called molten globule state [25,26]. On the other hand, there is sufficient evidence that the backbone conformation in the molten globule state is similar to that in the native state [27,28]. Therefore what we call the "native state" in our model is actually a molten globule with a preserved backbone fold.

Our calculations show that transition to the native state from disordered state where a chain does not have a definite conformation is thermodynamically a first-order phase transition. Moreover, such a character of the folding transition is a direct consequence of sequence selection. This is clearly seen in the phase diagram, Fig. 1, which shows that transitions which take place in random sequences, i.e., at $T > T_s^c$, are thermodynamically second order. This assertion is consistent with experimental data which show that formation of the molten globule state indeed occur as a first-order transition [29,30]. This should be contrasted with the behavior of random sequence [31] which has a broad transition from coil to the state which is similar to our globular disordered, or frozen state.

Even more important is the result which shows that for sequences generated in a certain range of selective temperatures the transition from the disordered state to the target conformation takes place at sufficiently high temperature, which is higher than T^{co}, the temperature at which the frozen state in random copolymer is stable. This has an important kinetic implication. It was shown in [14] that T^{co} plays a role of the glass transition temperature for random systems in a sense that relaxation to equilibrium becomes extremely slow at $T < T^{co}$ which prohibits folding at these temperatures. This fact inspired the introduction of "the principle of minimal frustrations" in [1,14]. The fact that for the designed sequences the native state is stable at $T > T^{co}$ makes this state also kinetically accessible which solves, for the chains with designed sequences, the folding problem.

ACKNOWLEDGMENTS

We thank Alexander Gutin, Alexander Grosberg, and Oleg Ptitsyn for interesting and useful discussions. This work was supported by the Packard Foundation.

[1] J.D. Bryngelson and P.G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).
[2] E.I. Shakhnovich and A.M. Gutin, J. Phys. A **22**, 1647 (1989).
[3] E.I. Shakhnovich and A.M. Gutin, Biophys. Chem. **34**, 187 (1989).
[4] E.I. Shakhnovich and A.M. Gutin, Europhys. Lett. **8**, 327 (1989); T. Garel and H. Orland, *ibid.* **6**, 307 (1988).
[5] E.I. Shakhnovich and A.M. Gutin, J. Phys. (Paris) **50**, 1843 (1989); T. Garel and H. Orland, Europhys. Lett. **6**, 597 (1988).
[6] G. Fredrickson and S. Milner, Phys. Rev. Lett. **67**, 835 (1991).
[7] S. Stepanow, M. Shulz, and J.-U. Sommer, Europhys. Lett. **19**, 273 (1992).
[8] C. Sfatos, A.M. Gutin, and E.I. Shakhnovich, Phys. Rev. E **48**, 465 (1993).
[9] A.M. Gutin and E.I. Shakhnovich, J. Chem. Phys. **98**, 8174 (1993).
[10] E.I. Shakhnovich and A.M. Gutin, Nature (London) **346**, 773 (1990).
[11] B. Derrida, Phys. Rev. B **24**, 2613 (1981).
[12] G. Kopper and H. Hilhorst, Europhys. Lett. **3**, 1213 (1987).
[13] E.I. Shakhnovich and A.M. Gutin, Europhys. Lett. **9**, 569 (1989).
[14] J.D. Bryngelson and P. Wolynes, J. Phys. Chem. **93**, 6902 (1989).
[15] A. Sali, E.I. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[16] Y. H. Taketomi and N. Go, Int. J. Pept. Protein Res. **7**, 445 (1975).
[17] N. Go and H. Abe, Biopolymers **20**, 991 (1981).
[18] E. Shakhnovich and A. Gutin, Proc. Natl. Acad. Sci. USA **90**, 7195 (1993).
[19] E. Shakhnovich and A. Gutin, Protein Eng. **6**, 793 (1993).
[20] S. Obukhov, J. Phys. A **19**, 3655 (1986).
[21] I. Lifshitz, A.Yu. Grosberg, and A.R. Khohlov, Rev. Mod. Phys. **50**, 683 (1978).
[22] K. Binder and A. Young, Rev. Mod. Phys. **58**, 801 (1986).
[23] M. Sasai and P. Wolynes, Phys. Rev. Lett. **65**, 2740 (1990).
[24] M. Sasai and P. Wolynes, Phys. Rev. A **46**, 7979 (1992).
[25] E.I. Shakhnovich and A.V. Finkelstein, Biopolymer **28**, 1667 (1989).
[26] O.B. Ptitsyn, *Protein Folding*, edited by T. Creighton (W.H. Freeman and Company, New York, 1992), Chap. 6, pp. 243–300.
[27] D. M. Harding and D. Woolfson, Biochemistry **30**, 3120 (1991).
[28] M. Peng and P. Kim, Biochemistry (to be published).
[29] V. Uversky, G. Semisotnov, R. Pain, and O. Ptitsyn, FEBS Lett. **314**, 89 (1994).
[30] A. Gittis, W. Stites, and E.E. Lattman, J. Mol. Biol. **232**, 718 (1993).
[31] A. Chaffotte *et al.*, Biochemistry **30**, 8067 (1991).

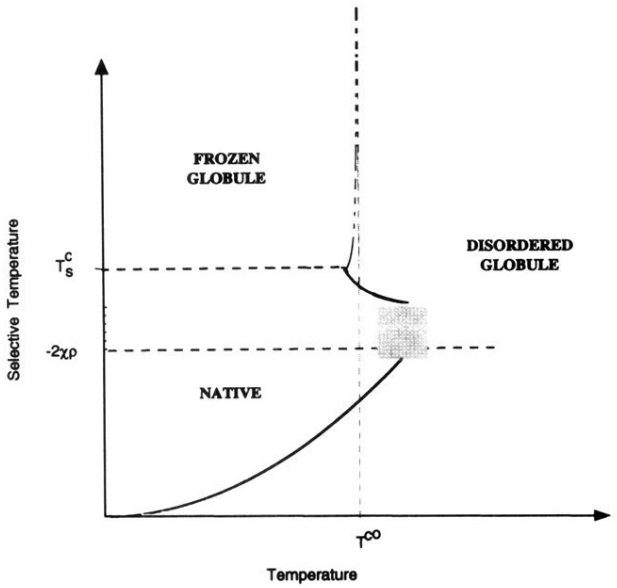

FIG. 1. The phase diagram shows the curves in the $T_{\rm sel}$, T space along which transitions from one phase to another take place. For sequences generated at $T_{\rm sel} > T_{\rm sel}^c$ the polymer goes into the frozen state on cooling in real space while for sequences generated at $T_{\rm sel} < T_{\rm sel}^c$ the disordered globule goes into the target conformation on cooling. The fat line corresponds to first-order transition. The shaded area denotes the region near the "ferromagnetic transition point" in sequence space where fluctuations are strong, and mean-field theory described here is inapplicable. However, for sequences generated at lower selective temperature the transition temperature to the target conformation decreases due to increased correlations in sequences.

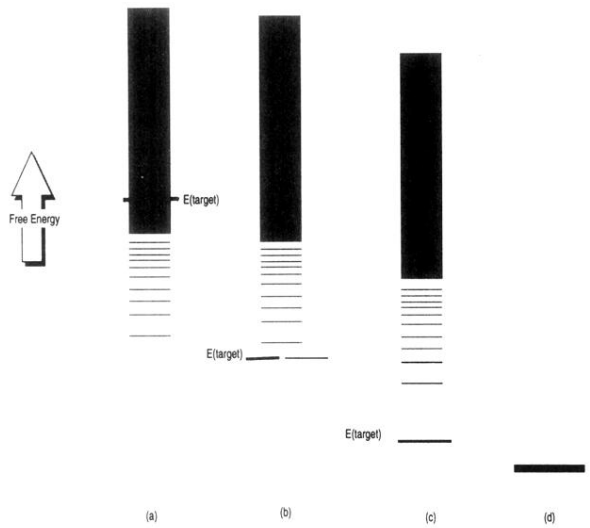

FIG. 2. (a) $T_{\mathbf{sel}} > T^{c}_{\mathbf{sel}}$, thus the energy of the target conformation is greater than that on the frozen states, hence on cooling the system goes into the frozen state. (b) $T_{\mathbf{sel}} = T^{c}_{\mathbf{sel}}$, this is transition selective temperature from the frozen state to the target state. (c) $T_{\mathbf{sel}} < T^{c}_{\mathbf{sel}}$ now, on cooling the system the system freezes into the target conformation (d) $T_{\mathbf{sel}} = 0$. The statistical properties are dominated by chains with thermodynamically large homopolymeric segments and thus have a highly degenerate ground state and the system does not freeze into the target conformation.