Phase transitions in heteropolymers with "secondary structure"

G. Z. Archontis and E. I. Shakhnovich*
Department of Chemistry, Harvard University, Cambridge, Massachusetts 02138
(Received 29 September 1993)

We study a model of a random heteropolymer with "secondary structure" at the level of mean field theory. The randomness in the polymer sequence is represented by a set of quenched disorder variables that describe the monomer-monomer interactions. The secondary structure is represented by a set of Ising-like thermodynamic variables that describe internal states of the monomers. The interactions between the monomers depend on the quenched disorder variables, on the thermodynamic, secondary-structure state variables, and also on the polymer configuration. We find that the system can exist in different phases that depend on the heterogeneity and average strength of the interactions, and on the polymer flexibility. At high temperatures the polymer interconverts freely between configurations without a stable secondary structure. At intermediate temperatures there is a transition to phases with one or two (coexisting) stable secondary-structure motifs and with a large number of thermodynamically important spatial configurations. At low enough temperatures (determined by the polymer flexibility) the polymer undergoes a freezing transition into phases with a unique spatial configuration and one or two stable secondary-structure motifs.

PACS number(s): 61.41.+e, 87.15.Da, 64.60.Cn, 64.60.Kw

I. INTRODUCTION

The protein folding problem is one of the most important questions in the field of molecular biology [1]. This problem consists in determining the factors that allow proteins to have a unique stable structure, and the factors that allow them to reach this structure without scanning the astronomically large number of possible conformations (the so called *Levinthal paradox* [2]). The first issue can be investigated with methods of equilibrium statistical mechanics, whereas the second requires a kinetic approach [3,4].

In recent years there has been substantial progress from a theoretical point of view, in elucidating the properties, that are *sufficient* to provide proteins with a unique structure. The progress was accomplished by modeling the proteins as random heteropolymers [5–8]. The properties of the random heteropolymers were studied using concepts from the theory of spin glasses [9].

The basic result of this approach is that these systems have an energy spectrum similar to the random energy model (REM) [10]. Each energy of the spectrum is associated with a particular conformation. The low energy part of the spectrum is discrete and corresponds to a few [of order $O(1)$] conformations that are completely different from each other. These few conformations become thermodynamically dominant below a temperature $T_{\rm tr}$ that depends on the heterogeneity of the interactions and the polymer flexibility. Thus crossing this temperature signals the freezing transition to a particular native structure.

The structural heterogeneity was introduced in [7,8] by a set of independent random variables B_{ij} that described the monomer-monomer interaction. In a set of other works the two-letter heteropolymer problem for a *sequence model* was also solved [11]. In this model the monomers were assigned quenched random variables σ_i ($= \pm 1$) and the monomer-monomer interaction for monomers in contact was equal to $\sigma_i\sigma_j$. The assumption of complete randomness in the polymer sequence (equivalently in the monomer-monomer interactions) simplifies the problem mathematically. In real proteins though it is rather certain that sequences are not completely random, but that they have been designed through evolution to satisfy certain structural and functional needs. This nonrandom character of sequences is obvious since (for example) there are many instances of periodic arrangements of polar-nonpolar residues in the protein primary sequence [12] that enhance the protein stability by creating a hydrophobic core and a hydrophilic exterior. However, since the nonrandom character of the protein primary sequences is expected to increase the stability and facilitate the kinetics of folding of proteins, the determination of the factors that enable folding to a unique structure in the *random* models actually solves a more stringent problem. Thus to reveal the necessary conditions for folding the modeling of proteins as random heteropolymers is satisfactory.

Even though these models are successful in reproducing protein characteristics such as the stable native structure, they constitute simplified representations that miss various features of the actual protein architecture. More specifically, these systems treat proteins as "beads on a string," where each monomer is a sphere that interacts with the other monomers via short-range forces. One important feature of proteins not represented by the models is their secondary structure. The majority of residues in

*Electronic address: eugene@diamond.harvard.edu

proteins (90%) exist in one of three secondary-structure motifs: 38% α-helical, 20% β-sheet, and 32% reverse turns [1]. The importance of secondary-structure formation for the folding and the stability of proteins has not been fully elucidated. It has been argued [13] that secondary-structure elements are involved in early stages of folding. Alternatively, it has been claimed on the basis of studies of short sequences in $d = 2$ lattices [14] that compactness induces formation of secondary structure, and that hydrogen bonding or amino acid propensities for a specific secondary structural motif are not needed. Various secondary-structure prediction schemes have been developed [15–17]. These schemes assume that the secondary structure is determined by the *local* sequence of short segments on the polypeptide chain, and therefore disregard the stabilization due to interactions of residues that are far away in sequence. This is probably the most important reason for the moderate predictive power of these methods.

In this work we introduce the concept of secondary-structure formation into the random heteropolymer model used previously [7,8] to study protein behavior. Our representation of secondary structure is introduced with a set of two "internal states" that the monomers can select. The criterion for occupation of these states is thermodynamic, i.e., the states are modeled by thermodynamic and not quenched variables. As we explain in Secs. II and VII, this representation gives to the secondary structure an energetic rather than a geometric significance. In these sections we explain though how a geometric interpretation of the secondary structure thus modeled can be achieved. We solve this model in the mean field approximation and we derive a rich phase diagram. At high temperature the polymer switches between conformations and secondary structures freely without any thermodynamic preference for a specific fold or a secondary-structure pattern. As the temperature is lowered, the polymer undergoes a ferromagnetic or spin-glass-like phase transition to a state with stable secondary structure. At lower temperatures (depending on the flexibility) the polymer freezes into a particular fold. The pattern of freezing into a fold is similar as for the case of simple heteropolymers [7,8,11], i.e., the thermodynamically dominant folds (below the freezing temperature) are completely different.

Our work is organized as follows. In Sec. II we present the model. In Sec. III we carry out the average over disorder and we present the order parameters that differentiate between the various phases. In Sec. IV we determine the freezing pattern in the space of folds and in Sec. V we use this pattern to determine the transition at the level of secondary structure. The reader who is not interested in the mathematical details of the calculations can skip Secs. IV and V. In Sec. VI we present and analyze the phase diagram. This is the basic result of our work. Finally, in Sec. VII we discuss our results and the limits of applicability of the model.

II. THE MODEL

Our purpose is to construct a model that represents a linear heteropolymer with secondary structure. The polymer is composed of N monomers that move in space subject to the constraint of linear arrangement on the polymer chain. Each monomer is assumed to have an excluded volume v. The monomers interact with each other with a potential that depends on their mutal separation. The secondary structure is introduced into the model as a property that affects the energy of interaction between the various monomers, and is modeled by a set of internal states that the monomers can adopt. These states could represent the α-helical vs random-coil (or β-sheet) monomer conformations observed in protein structures. Thus our model system is described by the Hamiltonian

$$\mathcal{H} = -\frac{1}{2}\sum_{i\neq j} D_{ij}\delta(\mathbf{r}_i - \mathbf{r}_j)\,\sigma_i\sigma_j + \frac{1}{2}B\sum_{i\neq j}\delta(\mathbf{r}_i - \mathbf{r}_j) + \frac{1}{6}C\sum_{i\neq j\neq k}\delta(\mathbf{r}_i - \mathbf{r}_j)\delta(\mathbf{r}_j - \mathbf{r}_k). \tag{2.1}$$

The first term incorporates the heteropolymeric interactions. These interactions depend on two kinds of *thermodynamic* variables, the spatial coordinates $\{\mathbf{r}_i\}$ and the set of internal states $\{\sigma_i\}$ that describe the secondary structure of the monomers, and a set of *quenched disorder* variables, the interaction strengths D_{ij}. Throughout this work we will use the following terminology. We will call *configuration* a set of the variables $\{\mathbf{r}_i, \sigma_i\}$. In contrast to this, a set of the r-space coordinates $\{\mathbf{r}_i\}$ will define a *fold*.

We chose the simplest short-range potential $U(\mathbf{r}_i - \mathbf{r}_j) = \delta(\mathbf{r}_i - \mathbf{r}_j)$ for the monomer-monomer interactions. This form of potential reproduces the essential behavior of all short-range two-body potentials [18]. The internal states of the monomers are represented by the set of variables $\{\sigma_i\}$. In this work we consider the case where the variables $\{\sigma_i\}$ take the values ± 1. This Ising-like representation corresponds to the situation where the monomers can exist in two distinct secondary-structure states.

The quenched disorder variables D_{ij} are assumed to be independent random variables that follow a Gaussian distribution

$$\mathcal{P}(D_{ij}) = \frac{1}{\sqrt{2\pi D^2}} e^{-\frac{(D_{ij}-D_0)^2}{2D^2}}, \tag{2.2}$$

with mean D_0 and variance D. The parameter D_0 controls the way the various secondary-structure elements interact with each other. For example, if $D_0 = 0$ any monomer i will have on the average the same number of positive and negative interaction strengths D_{ij} with the other monomers $j \neq i$, and it will stabilize on the average an equal number of similar and different secondary-structure states, respectively. On the other hand, a mean $D_0 > 0$ will lead to a larger number of positive parameters D_{ij}. In that case, a monomer will stabilize on the average neighboring monomers with similar secondary structure. In the following we will restrict ourselves to the case $D_0 \geq 0$.

The last two terms of the above Hamiltonian describe a background of homopolymeric interactions. The second term guarantees that the polymer will collapse to a

globular state, whereas the third term (for $C > 0$) prevents the monomers from collapsing to the same point. We will assume that $C > 0$ in the following.

The polymeric nature of the system is taken into account by including a term of the form

$$g(\mathbf{r}_{i+1} - \mathbf{r}_i) = \frac{1}{(2\pi a^2)^{3/2}} \exp\left(-\frac{(\mathbf{r}_{i+1} - \mathbf{r}_i)^2}{2a^2}\right) \quad (2.3)$$

in the polymer partition function. This term corresponds to the interactions between adjacent monomers, and expresses the fact that the average distance (Kuhn length) between adjacent monomers in the chain is equal to a.

Summarizing the characteristics of this model, our Hamiltonian corresponds to a system with heteropolymeric interactions superimposed on a background of homopolymeric interactions. Any two monomers interact when they are in contact, with an interaction strength that depends not only on the particular pair of monomers (i.e., on the sequence), but also on a set of internal states that characterize the monomers. One should note that a particular fold is defined merely by the set of coordinates $\{\mathbf{r}_i\}$. Thus the same fold can be associated with different sets of the $\{\sigma_i\}$ variables, or in other words with different "secondary structures." This is a consequence of the fact that the secondary structure, as introduced into this model, is a property that affects the energy of a configuration, but it is not directly related to its particular geometry. It is easy to understand the concept of a particular fold with more than one secondary structure, and also to attribute a geometric nature to this model of secondary structure, if one considers that any fold is defined up to the characteristic scale v of the monomer specific volume. The $\{\sigma_i\}$ variables could then account for additional details, such as the particular orientations of monomers that are in contact.

III. AVERAGE OVER DISORDER

The partition function for the Hamiltonian of Eq. (2.1) is given by the relation

$$Z(\{D_{ij}\}) = \sum_{\{\sigma_i\}} \int \prod_i d\mathbf{r}_i \prod_i g(\mathbf{r}_{i+1} - \mathbf{r}_i)\, e^{-\beta \mathcal{H}(\mathbf{r}_i, \{D_{ij}\})}. \quad (3.1)$$

Our objective is to calculate the free energy $\mathcal{F}(\{D_{ij}\}) = -KT \ln Z(\{D_{ij}\})$. To achieve this, we calculate instead the averaged over disorder (i.e., over sequences) free energy $\mathcal{F} = \langle -KT \ln Z(\{D_{ij}\})\rangle_{\rm av}$, where $\langle\,\rangle_{\rm av}$ denotes the average over the disorder variables D_{ij} with the weight given by Eq. (2.2). Since the free energy is a self-averaging quantity [9], $\mathcal{F}(\{D_{ij}\}) = \mathcal{F}$ in the thermodynamic limit. To calculate $\mathcal{F}$ we use the replica trick [9], which calls for averaging the nth moment of the partition function:

$$\mathcal{F} = \lim_{n\to 0} \frac{\langle Z(\{D_{ij}\})^n\rangle_{\rm av} - 1}{n}. \quad (3.2)$$

To carry out the average $\langle Z(\{D_{ij}\})^n\rangle_{\rm av}$ we introduce n replicas of the system and perform the integration over all sequences $\{D_{ij}\}$:

$$\begin{aligned}
\langle Z(\{D_{ij}\})^n\rangle_{\rm av} &= \sum_{\{\sigma_i^\alpha\}} \int \prod_{i,\alpha} d\mathbf{r}_i^\alpha\, g(\mathbf{r}_{i+1}^\alpha - \mathbf{r}_i^\alpha) \int \prod_{i<j} dD_{ij} \mathcal{P}(D_{ij}) \\
&\quad \times \exp\left\{ \frac{\beta}{2} \sum_{i\neq j} D_{ij} \sum_\alpha \delta(\mathbf{r}_i^\alpha - \mathbf{r}_j^\alpha) - \frac{\beta B}{2} \sum_{i\neq j} \sum_\alpha \delta(\mathbf{r}_i^\alpha - \mathbf{r}_j^\alpha) - \frac{\beta C}{6} \sum_{i\neq j\neq k} \sum_\alpha \delta(\mathbf{r}_i^\alpha - \mathbf{r}_j^\alpha)\delta(\mathbf{r}_j^\alpha - \mathbf{r}_k^\alpha) \right\} \\
&= \sum_{\{\sigma_i^\alpha\}} \int \prod_{i,\alpha} d\mathbf{r}_i^\alpha g(\mathbf{r}_{i+1}^\alpha - \mathbf{r}_i^\alpha) \\
&\quad \times \exp\left\{ -\beta \widetilde{B} \sum_\alpha \int d\mathbf{R} \left(\sum_i \delta(\mathbf{r}_i^\alpha - \mathbf{R}) \right)^2 - \frac{\beta C}{6} \sum_\alpha \int d\mathbf{R} \left(\sum_i \delta(\mathbf{r}_i^\alpha - \mathbf{R}) \right)^3 \right. \\
&\quad + \frac{(\beta D)^2}{2} \sum_{\alpha<\beta} \int d\mathbf{R}_1 d\mathbf{R}_2 \left(\sum_i \delta(\mathbf{r}_i^\alpha - \mathbf{R}_1)\delta(\mathbf{r}_i^\beta - \mathbf{R}_2)\, \sigma_i^\alpha \sigma_i^\beta \right)^2 \\
&\quad \left. + \frac{\beta D_0}{2} \sum_\alpha \int d\mathbf{R} \left(\sum_i \delta(\mathbf{r}_i^\alpha - \mathbf{R})\, \sigma_i^\alpha \right)^2 \right\}. \quad (3.3)
\end{aligned}$$

In arriving at Eq. (3.3) we ignored some additive constants and we denoted with $\widetilde{B}$ the quantity $\widetilde{B} = (B - C)/2 - (\beta D)^2/4$.

In Eq. (3.3) there appear three order parameters: (1) The polymer density $\rho_\alpha(\mathbf{R}) = \sum_i \delta(\mathbf{r}_i^\alpha - \mathbf{R})$, (2) The "secondary-structure" density $\eta_\alpha(\mathbf{R}) = \sum_i \delta(\mathbf{r}_i^\alpha - \mathbf{R})\, \sigma_i^\alpha$, and (3) the overlap parameter $\Psi_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) = \sum_i \delta(\mathbf{r}_i^\alpha - \mathbf{R}_1)\delta(\mathbf{r}_i^\beta - \mathbf{R}_2)\, \sigma_i^\alpha \sigma_i^\beta$. The order parameters $\eta_\alpha(\mathbf{R})$ and $\Psi_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2)$ satisfy the following normalization conditions:

$$\int d\mathbf{R}\,\eta_\alpha(\mathbf{R}) = \sum_i \sigma_i^\alpha,$$

$$\int d\mathbf{R}_1 d\mathbf{R}_2\,\Psi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2) = \sum_i \sigma_i^\alpha \sigma_i^\beta.$$

Thus $\eta_\alpha(\mathbf{R})$ provides information about the net secondary structure of the configuration α, whereas $\Psi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)$ is an indicator of the similarity in secondary structure of the two configurations α and β. As we will show in Sec. V, $\Psi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)$ is also related to the similarity of the folds that correspond to α and β.

In terms of these order parameters the average $\langle Z(\{D_{ij}\})^n\rangle_{\rm av}$ can be expressed as

$$\langle Z(\{D_{ij}\})^n\rangle_{\rm av} = \sum_{\{\sigma_i^\alpha\}} \int \prod_{i,\alpha} d\mathbf{r}_i^\alpha g(\mathbf{r}_{i+1}^\alpha - \mathbf{r}_i^\alpha) \exp\Big\{-\beta\widetilde{B}\sum_\alpha \int d\mathbf{R}\,\rho_\alpha^2(\mathbf{R}) - \frac{\beta C}{6}\sum_\alpha \int d\mathbf{R}\,\rho_\alpha^3(\mathbf{R}) + \frac{(\beta D)^2}{2}\sum_{\alpha<\beta}\int d\mathbf{R}_1 d\mathbf{R}_2\,\Psi_{\alpha\beta}^2(\mathbf{R}) + \frac{\beta D_0}{2}\sum_\alpha \int d\mathbf{R}\,\eta_\alpha^2(\mathbf{R})\Big\}. \tag{3.4}$$

The spatial density $\rho_\alpha(\mathbf{R})$ can be determined by minimizing the free energy with respect to $\rho_\alpha(\mathbf{R})$, subject to the constraint $\int d\mathbf{R}\,\rho_\alpha(\mathbf{R}) = N$ [8,19]. This minimization leads to the following value for the density:

$$\rho_\alpha(\mathbf{R}) = \frac{3}{C}\Big(\frac{C-B}{2} + \frac{(\beta D)^2}{4}\Big). \tag{3.5}$$

In the following we will assume that the polymer has a constant density with value given from Eq. (3.5) and we will designate this value with ρ. This assumption is justified, since in the globular state the density-density fluctuations vanish in the thermodynamic limit [19]. Using the constraint of constant density, we will omit from the following calculations the homopolymeric part of the Hamiltonian that depends on ρ.

To proceed further we perform a Hubbard-Stratonovitch transformation of Eq. (3.4) with respect to the order parameters $\eta_\alpha(\mathbf{R})$ and $\Psi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)$.

$$\langle Z(\{D_{ij}\})^n\rangle_{\rm av} = \int \mathcal{D}m_\alpha(\mathbf{R})\, e^{-\frac{\beta D_0}{2}\sum_\alpha \int d\mathbf{R} m_\alpha^2(\mathbf{R})} \int \mathcal{D}\varphi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2) e^{-\frac{(\beta D)^2}{2}\sum_{\alpha<\beta}\int d\mathbf{R}_1 d\mathbf{R}_2 \varphi_{\alpha\beta}^2(\mathbf{R}_1,\mathbf{R}_2)} \times \sum_{\{\sigma_i^\alpha\}}\int\prod_{i,\alpha} d\mathbf{r}_i^\alpha g(\mathbf{r}_{i+1}^\alpha - \mathbf{r}_i^\alpha)\exp\Big\{\beta D_0\sum_\alpha\int d\mathbf{R}\,m_\alpha(\mathbf{R})\eta_\alpha(\mathbf{R}) + (\beta D)^2\sum_{\alpha<\beta}\int d\mathbf{R}_1 d\mathbf{R}_2\,\varphi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)\Psi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)\Big\}. \tag{3.6}$$

This transformation introduces the auxiliary fields $m_\alpha(\mathbf{R})$ and $\varphi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)$. It is easy to check, with the help of Eq. (3.6), that these fields satisfy the saddle-point equations:

$$m_\alpha(\mathbf{R}) = \lim_{n\to 0}\frac{\mathrm{Tr}\,\eta_\alpha(\mathbf{R}) e^{L[\varphi_{\alpha\beta},m_\alpha]}}{\mathrm{Tr}\,e^{L[\varphi_{\alpha\beta},m_\alpha]}} \equiv \langle\eta_\alpha(\mathbf{R})\rangle, \tag{3.7}$$

$$\varphi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2) = \lim_{n\to 0}\frac{\mathrm{Tr}\,\Psi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2) e^{L[\varphi_{\alpha\beta},m_\alpha]}}{\mathrm{Tr}\,e^{L[\varphi_{\alpha\beta},m_\alpha]}} \equiv \langle\Psi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)\rangle, \tag{3.8}$$

where Tr denotes the operation $\mathrm{Tr}\equiv \int\prod_{i,\alpha} d\mathbf{r}_i^\alpha g(\mathbf{r}_{i+1}^\alpha - \mathbf{r}_i^\alpha)\sum_{\{\sigma_i^\alpha\}}$ and L denotes the function

$$L[\varphi_{\alpha\beta},m_\alpha] = \beta D_0\sum_\alpha\int d\mathbf{R}\,m_\alpha(\mathbf{R})\eta_\alpha(\mathbf{R}) + (\beta D)^2\sum_{\alpha<\beta}\int d\mathbf{R}_1 d\mathbf{R}_2\,\varphi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)\Psi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2). \tag{3.9}$$

As Eqs. (3.7) and (3.8) demonstrate, to solve the problem it is sufficient to evaluate the fields $\varphi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)$ and $m_\alpha(\mathbf{R})$. We thus need to perform the Tr that appears in Eq. (3.6). This will be the subject of the next two sections.

IV. PATTERN OF FREEZING IN THE SPACE OF FOLDS

We will first investigate the (thermodynamically important) overlap of the various configurations with respect to their spatial coordinates $\{\mathbf{r}_i^\alpha\}$. To achieve this we expand in Eq. (3.6) the exponential that contains the fields $\eta_\alpha(\mathbf{R})$ and $\Psi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)$, substitute the explicit form of $\eta_\alpha(\mathbf{R})$ and $\Psi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)$, and perform the summation over the $\{\sigma_i^\alpha\}$ variables. We then get for the terms up to $O(\varphi_{\alpha\beta}^4, m_\alpha^4)$

$$\langle Z(\{D_{ij}\})^n\rangle_{\rm av} = \int \mathcal{D}m_\alpha(\mathbf{R})\, e^{-\frac{\beta D_0}{2}\sum_\alpha \int d\mathbf{R} m_\alpha^2(\mathbf{R})} \int \prod_{i,\alpha} d\mathbf{r}_i^\alpha g(\mathbf{r}_{i+1}^\alpha - \mathbf{r}_i^\alpha)$$
$$\times \int \mathcal{D}\varphi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)\, e^{-\frac{(\beta D)^2}{2}\sum_{\alpha<\beta}\int d\mathbf{R}_1 d\mathbf{R}_2 \varphi_{\alpha\beta}^2(\mathbf{R}_1,\mathbf{R}_2)}$$
$$\times\Bigg\{1+\frac{(\beta D)^4}{4}\sum_{\{\alpha,\beta\}}\int \varphi_{\alpha\beta}^2(\mathbf{R}_1,\mathbf{R}_2)Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)+\frac{(\beta D_0)^2}{2}\rho\sum_\alpha\int m_\alpha^2(\mathbf{R})$$
$$+\frac{(\beta D)^6}{6}\sum_{\{\alpha,\beta,\gamma\}}\int \varphi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)\varphi_{\beta\gamma}(\mathbf{R}_2,\mathbf{R}_3)\varphi_{\gamma\alpha}(\mathbf{R}_3,\mathbf{R}_1)Q_{\alpha\beta\gamma}(\mathbf{R}_1,\mathbf{R}_2,\mathbf{R}_3)$$
$$+\frac{(\beta D)^2(\beta D_0)^2}{2}\sum_{\{\alpha,\beta\}}\int \varphi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)m_\alpha(\mathbf{R}_1)m_\beta(\mathbf{R}_2)$$
$$+\frac{(\beta D)^8}{48}\sum_{\{\alpha,\beta\}}\int \varphi_{\alpha\beta}^4(\mathbf{R}_1,\mathbf{R}_2)Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)$$
$$+\frac{(\beta D)^8}{24}\sum_{\{\alpha,\beta,\gamma,\delta\}}\int \varphi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)\varphi_{\beta\gamma}(\mathbf{R}_2,\mathbf{R}_3)$$
$$\times\varphi_{\gamma\delta}(\mathbf{R}_3,\mathbf{R}_4)\varphi_{\delta\alpha}(\mathbf{R}_4,\mathbf{R}_1)Q_{\alpha\beta\gamma\delta}(\mathbf{R}_1,\mathbf{R}_2,\mathbf{R}_3,\mathbf{R}_4)$$
$$+\frac{(\beta D)^8}{32}\sum_{\{\alpha,\beta,\gamma,\delta\}}\int \varphi_{\alpha\beta}^2(\mathbf{R}_1,\mathbf{R}_2)Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)\int \varphi_{\gamma\delta}^2(\mathbf{R}_3,\mathbf{R}_4)Q_{\gamma\delta}(\mathbf{R}_3,\mathbf{R}_4)$$
$$+\frac{(\beta D)^8}{32}\sum_{\{\alpha,\beta,\gamma\}}\int \varphi_{\alpha\beta}^2(\mathbf{R}_1,\mathbf{R}_2)\varphi_{\beta\gamma}^2(\mathbf{R}_2,\mathbf{R}_3)Q_{\alpha\beta\gamma}(\mathbf{R}_1,\mathbf{R}_2,\mathbf{R}_3)$$
$$+\frac{(\beta D)^4(\beta D_0)^2}{8}\rho\sum_{\{\alpha,\beta\}}\sum_\epsilon\int \varphi_{\alpha\beta}^2(\mathbf{R}_1,\mathbf{R}_2)Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)\int m_\epsilon^2(\mathbf{R})$$
$$+\frac{(\beta D)^4(\beta D_0)^2}{4}\sum_{\{\alpha,\beta,\gamma\}}\int \varphi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)\varphi_{\beta\gamma}(\mathbf{R}_2,\mathbf{R}_3)$$
$$\times Q_{\alpha\beta\gamma}(\mathbf{R}_1,\mathbf{R}_2,\mathbf{R}_3)m_\alpha(\mathbf{R}_1)m_\gamma(\mathbf{R}_3)$$
$$+\frac{(\beta D_0)^4}{24}\rho\sum_\alpha\int m_\alpha^4(\mathbf{R})+\frac{(\beta D_0)^4}{8}\rho^2\sum_{\{\alpha,\beta\}}\int m_\alpha^2(\mathbf{R})\int m_\beta^2(\mathbf{R}_2)+\cdots\Bigg\}, \tag{4.1}$$

where $\{\alpha,\beta,\ldots\}$ denotes summation over *distinct* indices. The integrals of the terms inside the braces $\{\}$ are over the variables $\{\mathbf{R}_i\}$ that appear in the respective integrands. In Eq. (4.1) we denoted with $Q_{\alpha\beta\ldots}(\mathbf{R}_1,\mathbf{R}_2,\ldots)$ the variables

$$Q_{\alpha\beta\ldots}(\mathbf{R}_1,\mathbf{R}_2,\ldots)=\sum_i \delta(\mathbf{r}_i^\alpha-\mathbf{R}_1)\delta(\mathbf{r}_i^\beta-\mathbf{R}_2)\cdots. \tag{4.2}$$

These variables are order parameters that have been used in previous studies of the simple heteropolymer [8,20] and the two-letter code [7,11], and they appear naturally in our problem after the summation over the $\{\sigma_i\}$ variables has been carried out.

The variables $Q_{\alpha\beta\ldots}(\mathbf{R}_1,\mathbf{R}_2,\ldots)$ incorporate information about the similarity in $\{r\}$ space of the replicas $\alpha,\beta,\ldots$. To see this, it suffices to consider (as an example) the overlap parameter $q_{\alpha\beta}$, defined as [8]

$$q_{\alpha\beta}\equiv\frac{1}{N}\sum_i\delta(\mathbf{r}_i^\alpha-\mathbf{r}_i^\beta)=\frac{1}{N}\int d\mathbf{R}\, Q_{\alpha\beta}(\mathbf{R},\mathbf{R}).$$

The parameter $q_{\alpha\beta}$ is related to the two-replica order parameter $Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)$ through the above equation. Thus, if $Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)$ is known, $q_{\alpha\beta}$ can be directly evaluated. It is straightforward to check that $q_{\alpha\beta}=0$ or 1 if the replicas α and β are, respectively, completely different or identical, and that it has an intermediate value in the case that α,β have some degree of similarity.

The overlap between replicas corresponds to the overlap between pure states [9,21]. In our case, a pure state is characterized by a set of conformations in the $\{r\}$ space and a set of conformations of the $\{\sigma_i\}$ variables. The variables $Q_{\alpha\beta\ldots}(\mathbf{R}_1,\mathbf{R}_2,\ldots)$ reveal the extent to which a set of pure states are similar in the $\{r\}$ space.

As it follows directly from Eq. (4.2), the order parameters $Q_{\alpha\beta\ldots}(\mathbf{R}_1,\mathbf{R}_2,\ldots)$ satisfy the normalization conditions:

$$\int d\mathbf{R}_2 d\mathbf{R}_3 \cdots Q_{\alpha\beta\ldots}(\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \ldots) = \rho_\alpha(\mathbf{R}_1),$$

$$\int d\mathbf{R}_1 d\mathbf{R}_2 d\mathbf{R}_3 \cdots Q_{\alpha\beta\ldots}(\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \ldots) = N.$$

In the case that the density ρ_α is constant in space, the above normalization conditions can be used to show [7,8,20] that the variables $Q_{\alpha\beta\ldots}$ depend on their arguments as $Q_{\alpha\beta\ldots}(\mathbf{R}_2 - \mathbf{R}_1, \mathbf{R}_3 - \mathbf{R}_1, \ldots)$ and that they obey the following scaling condition [7,8,20]:

$$Q_{\alpha\beta\gamma\ldots}(\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \ldots) = \frac{\rho}{R^{d(\nu-1)}} Q^{(1)}_{\alpha\beta\gamma\ldots}\left(\frac{\mathbf{R}_2 - \mathbf{R}_1}{R}, \frac{\mathbf{R}_3 - \mathbf{R}_1}{R}, \ldots\right). \quad (4.3)$$

In the above equation we introduced a characteristic length scale R associated with the difference between the positions $\{\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \ldots\}$ of the monomers in the replicas $\{\alpha, \beta, \gamma, \ldots\}$, respectively. We also denoted with d the dimensionality of the $\{r\}$ space and with ν the number of replicas in the definition of $Q_{\alpha\beta\gamma\ldots}$. The length scale R defines the extent to which the replicas $\{\alpha, \beta, \gamma, \ldots\}$ repeat each other in the $\{r\}$ space, and equivalently the scale up to which pure states are defined. Thus a pure state (in the $\{r\}$ space) can be viewed as a tube of characteristic size R. The function $Q^{(1)}_{\alpha\beta\gamma\ldots}$ that appears in Eq. (4.3) is defined [8] so that it satisfies the normalization condition $\int dxdy \cdots Q^{(1)}_{\alpha\beta\gamma\ldots}(x, y, \ldots) = 1$.

Since the variables $Q_{\alpha\beta\ldots}(\mathbf{R}_1, \mathbf{R}_2, \ldots)$ reveal the extent to which the pure states have similar folds, to determine the freezing pattern to particular folds we have to determine the values of $Q_{\alpha\beta\ldots}$ that maximize (for $n < 1$) the free energy. To accomplish this we need to express the free energy $\mathcal{F}$ as a function of $Q_{\alpha\beta\ldots}$. Our strategy is the following: Keeping the terms up to $O(m_\alpha^2, \varphi_{\alpha\beta}^2)$ and reexponentiating, Eq. (4.1) becomes

$$\langle Z(\{D_{ij}\})^n \rangle_{\text{av}} = e^{-\beta \mathcal{F}_1[\{m_\alpha\}]} \int \mathcal{D}Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2)\, e^{S[Q_{\alpha\beta}]} \times \int \mathcal{D}\varphi_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2)\, e^{-\frac{(\beta D)^2}{2} \sum_{\alpha<\beta} \int d\mathbf{R}_1 d\mathbf{R}_2 \left\{1 - (\beta D)^2 Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2)\right\} \varphi^2_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2)}, \quad (4.4)$$

where $\exp(-\beta \mathcal{F}_1[\{m_\alpha\}])$ denotes the contribution from the $\{m_\alpha\}$ integrations up to terms of order $O(m_\alpha^2)$. In Eq. (4.4) we switched integration variables from $\{\mathbf{r}_i^\alpha\}$ to $\{Q_{\alpha\beta}\}$ and we denoted with $S[\{Q_{\alpha\beta}\}]$ the quantity

$$S[\{Q_{\alpha\beta}\}] \equiv \ln \int \prod_{i,\alpha} d\mathbf{r}_i^\alpha g(\mathbf{r}_{i+1}^\alpha - \mathbf{r}_i^\alpha)\, \delta\left(Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) - \sum_i \delta(\mathbf{r}_i^\alpha - \mathbf{R}_1)\delta(\mathbf{r}_i^\beta - \mathbf{R}_2)\right). \quad (4.5)$$

It is clear that $\exp S$ is the number of folds in the n-replica $\{r\}$ space that correspond to two-replica overlaps equal to $Q_{\alpha\beta}$. Thus S is the $\{r\}$-space configurational entropy when each replica is choosing spatial folds from a tube of diameter R, with R the characteristic scale of $Q_{\alpha\beta}$. When $R \to \infty$ this tube extends over all space, i.e., each replica is allowed to sample all possible folds (with density ρ). In the opposite limit, when $R \to v^{1/3}$ each replica samples folds that differ only at scales smaller than the "resolution" of our model $v^{1/3}$.

The integral over the variables $\{\varphi_{\alpha\beta}\}$ in Eq. (4.4) is Gaussian to this order and can be performed exactly, leading to the result

$$\langle Z(\{D_{ij}\})^n \rangle_{\text{av}} = e^{-\beta \mathcal{F}_1[\{m_\alpha\}]} \int \mathcal{D}Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) e^{S[Q_{\alpha\beta}]} \exp\left\{-\frac{1}{2} \sum_{\alpha<\beta} \int d\mathbf{R}_1 d\mathbf{R}_2 \ln[1 - (\beta D)^2 Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2)]\right\}$$
$$= e^{-\beta \mathcal{F}_1[\{m_\alpha\}]} \int \mathcal{D}Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) \exp\left\{S[Q_{\alpha\beta}] + \frac{(\beta D)^4}{4} \sum_{\alpha<\beta} \int Q^2_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) + O((\beta D)^6 Q^4_{\alpha\beta})\right\}. \quad (4.6)$$

In arriving at the rightmost term of Eq. (4.6) we omitted some constants and we expanded the logarithm appearing in the second term. The higher-order terms that appear in the expansion of the right-hand side of Eq. (4.6) should be combined with the higher-order contributions from the integration of the terms of Eq. (4.1), that were omitted in arriving at Eq. (4.4), to evaluate correctly the free energy $\mathcal{F}$. However, all these terms do not affect the freezing pattern into specific folds, as we will show in the Appendix. Therefore to determine the freezing pattern (that is, the form of $Q_{\alpha\beta}$) we can disregard these terms for now.

In Eq. (4.6) the free energy $\mathcal{F}$ has been expressed as a function of the order parameters $Q_{\alpha\beta}$. A subsequent maximization of $\mathcal{F}$ with respect to $Q_{\alpha\beta}$ will determine the possible overlaps between the folds of different replicas. To proceed, it is convenient to maximize the free energy with respect to the characteristic length scale R instead of $Q_{\alpha\beta}$, as it was done in [7,8,11,20]. The entropy S has been shown in [8] to scale as $\sim -1/R^2$. This is intuitively expected, since this scaling law corresponds to the configurational entropy of a polymer in a tube with

diameter R [22]. On the other hand, the $O(Q_{\alpha\beta}^2)$ term in Eq. (4.6) scales as $\sim 1/R^3$, as it follows with the help of Eq. (4.3). Thus a maximization (for $n < 1$) of the free energy with respect to R leads to the values $R = \infty$ and $R = v^{1/3}$. The former value corresponds to the case where $Q_{\alpha\beta} = 0$, i.e., the replicas (the pure states) are associated with completely different folds. The latter value corresponds to the case where the replicas repeat themselves inside a tube of microscopic scale, i.e., the pure states are identical frozen folds. The resulting free energy is depicted as a function of $1/R$ in Fig. 1.

The contribution of higher-order terms can be calculated in a similar way (by evaluating moments of the above Gaussian integral). As an example, we present the calculation for the other terms of Eq. (4.1) in the Appendix. All these terms result in contributions to the free energy that scale as higher (≥ 3) powers of $1/R$. These terms affect the detailed appearance of the free energy $\mathcal{F}$ at microscopic scales, but they do not change the qualitative shape of the free energy curve depicted in Fig. 1. The values of $\mathcal{F}$ at $R \to \infty$ and $R \to v^{1/3}$ will still be separated by a free energy barrier, signifying that the freezing pattern corresponds to either $R \to \infty$ or $R \to v^{1/3}$.

The above freezing pattern suggests that the replicas can be divided into groups. Each group consists of replicas with identical folds, and the replicas that belong to different groups correspond to completely different folds. The form of the order parameters $Q_{\alpha\beta\gamma\ldots}$ is given by the relation

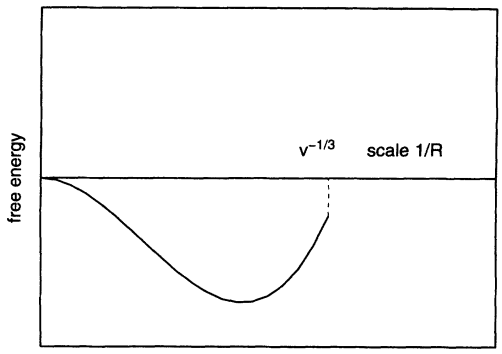

FIG. 1. Free energy $\mathcal{F}$ as a function of the $\{r\}$-space length scale $1/R$, plotted for the case $n < 1$. The free energy is maximized for $R \to \infty$ or $R = v^{1/3}$. In the former case a replica samples the entire $\{r\}$ space (no freezing), whereas in the latter case it is localized to the microscopic scale $v^{1/3}$ (freezing).

$$Q_{\alpha\beta\gamma\ldots}(\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \ldots) = \begin{cases} \rho\delta(\mathbf{R}_1 - \mathbf{R}_2)\delta(\mathbf{R}_1 - \mathbf{R}_3)\cdots & \text{for } \alpha, \beta, \gamma, \cdots \text{ in the same group} \\ 0 & \text{otherwise .} \end{cases} \tag{4.7}$$

The δ functions that appear in Eq. (4.7) are more precisely Dirac δ-function-*like* functions that are nonzero when their argument is of the order of $v^{1/3}$ [e.g., they could be defined as $\delta(x) = v^{-1}$ for $x \leq v^{1/3}$].

The Gaussian integral of Eq. (4.4) diverges in the case $(\beta D)^2 Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) > 1$. This divergence signals the existence of a phase transition that causes the variables $\{\varphi_{\alpha\beta}\}$ to acquire a fixed value. As we will demonstrate in the next section, this phase transition is associated with the behavior of both the $\{\mathbf{r}_i^\alpha\}$ and $\{\sigma_i^\alpha\}$ variables. One can see with the help of the above condition that the phase transition in $\{\varphi_{\alpha\beta}\}$ cannot occur as long as the freezing in the $\{r\}$ space has not taken place, since in this case $Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) = N/V^2 \to 0$ [8] in the thermodynamic limit.

V. FREEZING AT THE LEVEL OF SECONDARY STRUCTURE

As we established in the preceding section, the variables $Q_{\alpha\beta\ldots}$ can have two possible values given by Eq. (4.7). Thus we can arrange $Q_{\alpha\beta\ldots}$ in a Parisi-type matrix with these two values separated by a breakpoint x_0. This has the meaning that the n replicas are divided into n/x_0 groups, with x_0 replicas of identical $\{r_i^\alpha\}$ coordinates in each group.

Since we solve the problem at the level of mean field, we will set $m_\alpha(\mathbf{R}) =$const and $\varphi_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) = \varphi_{\alpha\beta}(\mathbf{R}_1 - \mathbf{R}_2)$. Substituting the form of $Q_{\alpha\beta}$ into Eq. (4.1), we find

$$\langle Z(\{D_{ij}\})^n \rangle_{\text{av}} = e^{S[Q_{\alpha\beta}]} \int \mathcal{D}m_\alpha \int \mathcal{D}\varphi_{\alpha\beta}(\mathbf{R}) \exp\Bigg\{ -\frac{\beta D_0 V}{2} \sum_{\alpha<\beta} m_\alpha^2 - \frac{(\beta D)^2 V}{2} \sum_{\alpha<\beta} \int d\mathbf{R}\, \varphi_{\alpha\beta}^2(\mathbf{R}) + N \sum_g \ln\Bigg[\sum_{\{\sigma_\alpha, \alpha\in g\}} \exp\Big((\beta D)^2 \sum_{\alpha<\beta\in g} \varphi_{\alpha\beta}(v^{1/3})\, \sigma_\alpha\sigma_\beta + \beta D_0 \sum_{\alpha\in g} m_\alpha\sigma_\alpha \Big) \Bigg] \Bigg\}, \tag{5.1}$$

where V is the volume of the system and $\{g\}$ denotes summation over all groups g of different folds. Equation (5.1) can be immediately derived from Eq. (3.6). To see this it is sufficient to notice via Eq. (3.8) that the separation of replicas into groups of different folds and the mean field functional form $\varphi_{\alpha\beta}(\mathbf{R}_1 - \mathbf{R}_2)$ are equivalent to setting $\Psi_{\alpha\beta}(\mathbf{R}_1 - \mathbf{R}_2) = \rho\Delta(\mathbf{R}_1 - \mathbf{R}_2) \sum_i \sigma_i^\alpha \sigma_i^\beta / N$ for α, β in different folds and 0 otherwise. In Eq. (5.1) we

denoted with $S[Q_{\alpha\beta}]$ the entropy that corresponds to the values of $Q_{\alpha\beta}$, as they are given by Eq. (4.7).

A minimization over $\{\varphi_{\alpha\beta}\}$ of the exponent in the integrand of Eq. (5.1) shows that $\varphi_{\alpha\beta}(\mathbf{R}) = 0$ for α, β in different folds and $\varphi_{\alpha\beta}(\mathbf{R}) = 0$, $R > v^{1/3}$ for α, β in the same fold. Thus, upon substituting $\varphi_{\alpha\beta}(\mathbf{R}) = \rho/v\,\widetilde{\varphi}_{\alpha\beta}$ for $R \leq v^{1/3}$ and α, β in the same fold, and 0 otherwise, and rescaling $m_\alpha \to \rho\widetilde{m}_\alpha$ we arrive at the following expression for the free energy:

$$\mathcal{F} = \lim_{n\to 0} \frac{1}{n} N \sum_g \left\{ \frac{\kappa}{2} \sum_{\alpha\in g} \widetilde{m}_\alpha^2 + \frac{\lambda}{2} \sum_{\alpha<\beta\in g} \widetilde{\varphi}_{\alpha\beta}^2 - \ln \sum_{\{\sigma_\alpha, \alpha\in g\}} \exp K_g[\widetilde{\varphi}_{\alpha\beta}, \widetilde{m}_\alpha] \right\}, \tag{5.2}$$

where the function K_g is given by the relation

$$K_g[\widetilde{m}_\alpha, \widetilde{\varphi}_{\alpha\beta}] = \lambda \sum_{\alpha<\beta\in g} \widetilde{\varphi}_{\alpha\beta}\sigma_\alpha\sigma_\beta + \kappa \sum_{\alpha\in g} \widetilde{m}_\alpha\sigma_\alpha, \tag{5.3}$$

with $\kappa = \beta D_0 \rho$ and $\lambda = (\beta D)^2 \rho/v$.

The rescaled variables $\widetilde{m}_\alpha$ and $\widetilde{\varphi}_{\alpha\beta}$ satisfy the normalization conditions $\int d\mathbf{R}\, m_\alpha(\mathbf{R}) = N\widetilde{m}_\alpha = N\langle\sigma_\alpha\rangle$ and $\int d\mathbf{R}_1 d\mathbf{R}_2\, \varphi_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) = N\widetilde{\varphi}_{\alpha\beta} = N\langle\sigma_\alpha\sigma_\beta\rangle$, where the average $\langle\,\rangle$ involves a summation over $\{\sigma_\alpha\}$, with a weight function $\exp K_g$. Then, from the mean field theory of spin glasses [9] we deduce that $\widetilde{m}_\alpha = \langle\langle\sigma_i\rangle_{\text{th}}\rangle_{\text{av}}$ and $\widetilde{\varphi}_{\alpha\beta} = \langle\langle\sigma_i\rangle_{\text{th}}^2\rangle_{\text{av}}$. For the sake of brevity in the following we will drop the tilde sign and denote the *rescaled* variables as m_α and $\varphi_{\alpha\beta}$.

A. Replica symmetric solution inside a folding group

As it is deduced from Eq. (5.1) and the following discussion, the overlap of $\{\sigma_i\}$ for replicas that correspond to *different folds* is zero. This is expected, since these folds have completely different contacts. For replicas that belong to the *same* fold, there might be a range of parameters for which $\varphi_{\alpha\beta} \neq 0$. Thus in terms of *all* replicas the nontrivial solution will always be replica-symmetry breaking. For replicas inside the same folding group one can seek solutions of $\varphi_{\alpha\beta}$ that are replica symmetric, or one can break the replica symmetry in a Parisi-type scheme. In the first case, if one sets $\varphi_{\alpha\beta} = \varphi, \forall\alpha, \beta \in g$ and $m_\alpha = m$, one arrives at the following equation for the (intensive) free energy:

$$\beta f = -\ln 2 + \frac{\lambda}{4}(x_0 - 1)\varphi^2 + \frac{\lambda}{2}\varphi + \kappa m^2 - \frac{1}{x_0} \ln \frac{1}{\sqrt{2\pi}} \int dp\, e^{-\frac{p^2}{2}} \cosh^{x_0}(\beta\mathcal{H}_{\text{eff}}) - \frac{1}{N} \lim_{n\to 0} \frac{1}{n} S[Q_{\alpha\beta}], \tag{5.4}$$

with $\beta\mathcal{H}_{\text{eff}} = \kappa m + \sqrt{\lambda\varphi}p$. The variables φ and m are determined by the following self-consistent equations:

$$m = \frac{\int dp\, e^{-p^2/2} \cosh^{x_0}(\beta\mathcal{H}_{\text{eff}}) \tanh(\beta\mathcal{H}_{\text{eff}})}{\int dp\, e^{-p^2/2} \cosh^{x_0}(\beta\mathcal{H}_{\text{eff}})}, \tag{5.5}$$

$$\varphi = \frac{\int dp\, e^{-p^2/2} \cosh^{x_0}(\beta\mathcal{H}_{\text{eff}}) \tanh^2(\beta\mathcal{H}_{\text{eff}})}{\int dp\, e^{-p^2/2} \cosh^{x_0}(\beta\mathcal{H}_{\text{eff}})}. \tag{5.6}$$

Equations (5.5) and (5.6) are different from the replica-symmetric equations for the Sherrington-Kirkpatrick (SK) model [9,23], due to the nontrivial dependence of φ and m on the parameter x_0. In order to determine x_0 one needs a third equation, in addition to Eqs. (5.5) and (5.6). This equation is deduced by maximization of the free energy $\mathcal{F}$ with respect to x_0. To perform this maximization one needs to know the form of the entropy $S[Q_{\alpha\beta}]$. This form was derived in [8], where it was shown that

$$S[Q_{\alpha\beta}] = N\frac{n}{x_0}(x_0 - 1)\ln\frac{v}{a^3}. \tag{5.7}$$

The form of $S[Q_{\alpha\beta}]$ as given by Eq. (5.7) can be easily elucidated, if one notices that $S[Q_{\alpha\beta}]$ is also equal to the *change* in the entropy due to the freezing. This holds because the entropy for the random coil is $S = 0$ due to the normalization of the functions $g(\mathbf{r}_{i+1} - \mathbf{r}_i)$. It is easy to check then that the form of $S[Q_{\alpha\beta}]$ as given by Eq. (5.6) corresponds to the above described freezing pattern. Selecting one replica from a group as a reference, any monomer of a second replica in the same group has to be positioned in a space of volume v, due to the localization scale, instead from the space a^3 that would correspond to a random-coil segment of Kuhn length a. Thus for any monomer of the second replica the entropy loss is $\ln v/a^3$. Taking into account the fact that there are $x_0 - 1$ members in each group (in addition to the reference replica) and n/x_0 groups, we arrive at Eq. (5.7).

Using Eqs. (5.4) and (5.7), one can maximize the free energy with respect to x_0 and deduce the following equation:

$$-\frac{\lambda}{4}\varphi^2 - \frac{1}{x_0^2} \ln \frac{1}{\sqrt{2\pi}} \int dp\, e^{-\frac{p^2}{2}} \cosh^{x_0}(\beta\mathcal{H}_{\text{eff}}) + \frac{1}{x_0} \times \frac{\int dp\, e^{-\frac{p^2}{2}} \cosh^{x_0}(\beta\mathcal{H}_{\text{eff}}) \ln\cosh(\beta\mathcal{H}_{\text{eff}})}{\int dp\, e^{-\frac{p^2}{2}} \cosh^{x_0}(\beta\mathcal{H}_{\text{eff}})} = -\frac{1}{x_0^2} \ln\frac{v}{a^3}. \tag{5.8}$$

Equations (5.5), (5.6), and (5.8) determine the variables m, φ and x_0 as functions of the disorder parameters D_0, D, the temperature T, and the parameter $\ln v/a^3$. This last parameter is related to the flexibility of the polymer, since as we explained above, a^3/v is a measure of the number of folds lost per monomer upon freezing into a particular fold. By inspection of Eqs. (5.5), (5.6), and (5.8) it is seen that the values of m and φ depend crucially on the value of the parameter x_0 (that is, on the freezing to a particular fold) and vice versa. It is this interdependence of the secondary-structure and spatial-folding order parameters that makes this model different

from the SK model. A numerical solution of the above equations allows the construction of a phase diagram that we describe and analyze in Sec. VI.

B. Replica-symmetry breaking inside folds

The form of the free energy $\mathcal{F}$, as given by Eqs. (5.2) and (5.3), is similar (in terms of the $\{\sigma\}$ variables) to the SK free energy. As it has been established in various works [24–26], the correct solution for a specific range of the parameters T, D_0, and D involves replica-symmetry breaking (RSB) with respect to the variables $\varphi_{\alpha\beta}$. This region lies below the de Almeida–Thouless (AT) line [9,24]. The solution described in Sec. V A involves RSB between different folds, because $\varphi_{\alpha\beta}=0$ for α,β in different folds, whereas $\varphi_{\alpha\beta}\neq 0$ (possibly) for α,β in the same fold. We can extend the RSB inside each fold, if we assume in the spirit of Parisi that $\varphi_{\alpha\beta}$ are functions of a continuous variable x, $\varphi_{\alpha\beta}=\varphi(x)$. In this case we set $\varphi(x)=0$ for $x<x_0$ (when $n<0$). For simplicity, we examine the case $D_0=0$. The results for $D_0\neq 0$ can be deduced from the $D_0=0$ solution at nonzero field, as it has been explained in [9,27]. Upon expanding the trace in Eq. (5.2) we find the following form for the free energy, up to terms of order $O(\varphi^4)$:

$$\beta f=-\frac{1}{x_0}(x_0-1)\ln\frac{v}{a^3}+\frac{\lambda}{4}(1-\lambda)\int_1^{x_0}dx\,\varphi^2(x)$$
$$-\frac{\lambda^3}{6}\int_1^{x_0}dx\left\{(x_0-x)\,\varphi(x)^3+\varphi(x)\int_x^{x_0}dy\,\varphi^2(y)+2\varphi^2(x)\int_1^{x}dy\,\varphi(y)\right\}-\frac{\lambda^4}{12}\int_1^{x_0}dx\,\varphi^4(x). \tag{5.9}$$

In deriving Eq. (5.9) we kept among the $O(\varphi^4)$ terms the one responsible for the RSB [28,29]. In the case of the SK model, retaining the other terms has been shown [30] to lead to a different functional form (i.e., nonlinear) for the function $\varphi(x)$ that maximizes the free energy, but close to the transition [where $\varphi(x)\to 0$] the two forms coincide. Parisi [31] has used a coefficient $-1/4$ for the quartic term. We will use the $-1/12$ coefficient that appears in [28,29].

Variation of the free energy with respect to φ gives

$$\frac{\lambda}{2}(1-\lambda)\varphi-\frac{\lambda^3}{6}\left\{3(x_0-x)\,\varphi^2+3\int_x^{x_0}dy\,\varphi^2(y)\right.$$
$$\left.+6\varphi(x)\int_1^{x}dy\,\varphi(y)\right\}-\frac{\lambda^4}{3}\varphi^3=0 \tag{5.10}$$

and differentiations with respect to x lead to the result

$$\varphi'(x)=0 \quad \text{or} \quad \varphi(x)=\frac{x-x_0}{2\lambda}. \tag{5.11}$$

If we assume that there are two breakpoints x_1 and x_2, with $0\le x_0\le x_1\le x_2\le 1$ between which $\varphi(x)$ assumes the linear form, i.e.

$$2\lambda\varphi(x)=\begin{cases}x_1-x_0, & x_0\le x\le x_1\\ x-x_0, & x_1\le x\le x_2\\ x_2-x_0, & x_2\le x\end{cases} \tag{5.12}$$

we find for x_1 and x_2

$$x_1=x_0 \quad \text{and} \quad x_2=1-\sqrt{(1-x_0)^2+\frac{2}{\lambda}(1-\lambda)}. \tag{5.13}$$

Since $x_2\ge x_0$ it follows from Eq. (5.13) that $\lambda\ge 1$. The value $\lambda=1$ is critical because for this value it follows that $x_2=x_1=x_0$, and the RSB inside each fold disappears. For $\lambda<1$ the solution for φ is given by the replica-symmetric (inside each fold) Eqs. (5.5) and (5.6). The critical temperature below which the RSB solutions given by Eq. (5.12) are correct is $T_{\rm tr}=D\sqrt{\rho/v}$, as it follows from the definition of λ.

The breakpoint x_0 has to be found by maximization of the free energy. Using Eq. (5.9) and keeping terms of order $O(\lambda^2)$, we find the following equation for x_0:

$$-\frac{1}{x_0^2}\ln\frac{v}{a^3}=\frac{\lambda-1}{16\lambda}(x_2-x_0)^2. \tag{5.14}$$

Equation (5.14) shows that the position of x_0 depends (in addition to λ) on the flexibility parameter $\ln v/a^3$. Since $\ln v/a^3\le 0$, for a physically meaningful solution $x_0>0$ to exist the condition $\lambda>1$ must be satisfied. This is consistent with the allowed values for λ that we arrived at before. For $\lambda\to 1^+$, where $x_2\to x_0$ Eq. (5.14) is satisfied only if $\ln v/a^3\to 0^-$. In this case, one can substitute the value of x_2 from Eq. (5.13) into Eq. (5.14) and show that $x_0<1$. Thus for $\ln v/a^3\simeq 0^-$ the *freezing* temperature $T_{\rm fr}$ at which x_0 departs from the value $x_0=1$ is $T_{\rm fr}\simeq T_{\rm tr}^-$. The value $T_{\rm fr}=T_{\rm tr}$ is the *maximum* critical temperature, below which freezing into particular folds can occur. If $\ln v/a^3\ll 0$, the freezing temperature $T_{\rm fr}$ drops considerably below $T_{\rm tr}$. This happens because freezing is opposed by the accompanying loss in entropy. It is important to note that freezing into a particular fold is possible only *after* the $\{\sigma_i^\alpha\}$ variables are able to undergo a phase transition. If this were not the case, our model would represent the unrealistic situation where the polymer (for a range of temperatures) would freeze into a definite fold, and the monomers would interconvert freely between secondary-structure states, with a zero average secondary structure.

For $T_{\rm fr}<<T_{\rm tr}$ Eq. (5.14) cannot be satisfied. This is a consequence of the fact that the expansion of Eq. (5.2) breaks down. This expansion implied that $\varphi(x)\to 0$, which holds only at $T\to T_{\rm tr}$ ($\lambda\to 1$).

The function $\varphi(x)$ is presented in Fig. 2 for the case

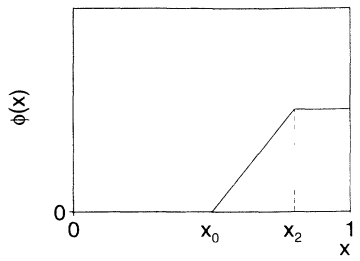

FIG. 2. The function φ that maximizes the free energy $\mathcal{F}$ in the RSB inside folds case, plotted for $T < T_{\rm fr} < T_{\rm tr}$. The linear functional form implies that $x_0 \to 1$. For clarity we plotted x_0, x_2 away from 1 and φ large.

$T < T_{\rm fr} < T_{\rm tr}$ and $T \to T_{\rm tr}$. For $x < x_0$, $\varphi(x) = 0$. This holds, because in this case the replicas belong to different groups (i.e., different folds), and therefore they are characterized by completely different contacts. Thus the $\{\sigma^\alpha\}$ overlap for such folds will be zero. The overlap departs from its zero value for $x > x_0$, and reaches a maximum value that depends on temperature. For $T \to 0$ this value is $\varphi \to 1$ [9,32].

In the case $x_0 = 1$ (i.e., for $T > T_{\rm fr}$) the function $\varphi(x) = 0$. This means that the spin-glass transition can be observed (in terms of the order parameter φ) only if freezing to a few (thermodynamically dominant) folds has occurred. On the other hand, we showed that the maximum critical temperature below which φ can be different from zero is $T_{\rm tr}$. This means that the variables $\{\sigma^\alpha\}$ can undergo a spin-glass transition for $T \leq T_{\rm tr}$. This transition does not show up in terms of the φ variables for $T_{\rm fr} < T \leq T_{\rm tr}$, because in this temperature range there exists a very large number of thermodynamically important folds with zero overlap. Since $\varphi = \sum_{\alpha,\beta} w_\alpha w_\beta \varphi_{\alpha\beta}$ with w_α the thermodynamic weight of the pure state α, it follows that $\varphi = 0$ if freezing to a few (thermodynamically dominant) folds has not occurred.

For $D_0 \neq 0$ an analysis by Toulouse for the SK model [9,27] has shown that m departs from the value $m = 0$ for $D_0 \geq D/\sqrt{\rho v}$. We discuss this case in more detail in the next section.

VI. PHASE DIAGRAM

Using the results of the preceding section we can construct the phase diagram for the random heteropolymer with dynamical variables $\sigma_i = \pm 1$. This diagram depends on the homopolymeric parameters B, C, the disorder parameters D_0, D, the temperature T, and the flexibility parameter $\ln v/a^3$. We will assume that the parameters B and C are chosen so that the polymer will exist in a globular phase [i.e., ρ as given by Eq. (3.6) satisfies $\rho > 0$]. Then, it is convenient to fix $\ln v/a^3$ at a nontrivial value ($\neq 0$) and construct the phase diagram as a function of the parameters $D_0/[D\sqrt{1/(\rho v)}]$ and $T/(D\sqrt{\rho/v})$.

The phase diagram derived in this way is presented in Fig. 3. In this figure we associated the parameters $D_0/[D\sqrt{1/(\rho v)}]$ and $T/(D\sqrt{\rho/v})$ with the x and y axis, respectively. In this way the comparison with the SK phase diagram can be directly made. The parameter $\ln v/a^3$ can be associated with a third direction (i.e., vertical to the plane). Having fixed $\ln v/a^3$ we then look at a particular slice of the three-dimensional (3D) phase diagram.

The various phases are defined by the values of m, φ, and x_0. *Note that with m and φ we refer to the rescaled variables,* as these were defined in Sec. V. The values of these variables can be determined either by assuming that inside each fold φ is replica symmetric and using Eqs. (5.5), (5.6), and (5.8), or by assuming that φ obeys RSB inside each fold. We will choose the latter solution for the region below the AT line. As it has been demonstrated previously for the SK model [24–26], in this region a RSB solution lowers the free energy. Our $\sigma = \pm 1$

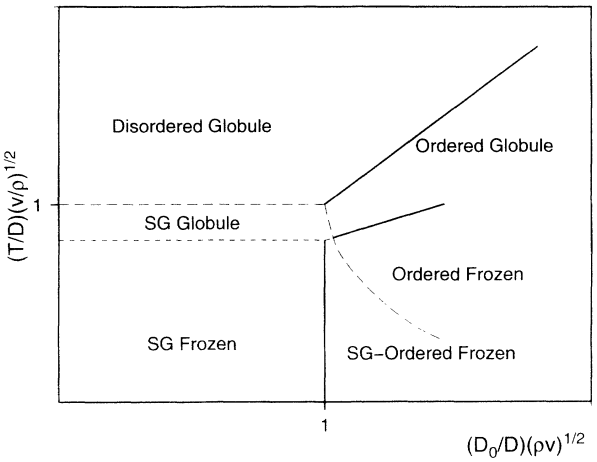

FIG. 3. Phase diagram for a heteropolymer with secondary structure. The term globule refers to a phase in which the polymer interconverts between folds and the term frozen to a phase with a few [of order $O(1)$] folds. In the disorderd globule phase the polymer alternates between folds and the monomers switch freely between secondary-structure states, without any thermodynamic preference for a specific fold or secondary-structure motif. In the ordered globule phase the polymer has a predominating secondary structure, and in the ordered frozen phase the polymer has a unique native structure and a predominating secondary structure. In these two ordered phases *any* monomer spends most of the time in the predominating state, and it switches between states due to thermal fluctuations. In the SG phases the thermodynamic preference for a secondary-structure state changes from monomer to monomer, with *both* states observed along a fold. In the SG-ordered frozen phase one state is in excess, whereas in the SG frozen phase both states are present to the same extent. In the SG globule phase the polymer interconverts between folds, with a secondary structure that changes in a fold-dependent manner. The *folding* line (below which freezing occurs) is depicted as short-dashed inside and continuous outside the SG region. The long-dashed line is the AT line. The secondary-structure-related order parameter φ assumes replica-symmetric solutions (for the replicas in the same folding group) above the AT line.

heteropolymer is related to the SK model, as is apparent from the above analysis. In Fig. 3 we plotted the AT line as a long-dashed line. The AT line was determined numerically by solving Eqs. (5.5) and (5.6) with $x_0 = 0$ (i.e., the ordinary SK equations). Above the AT line we solved Eqs. (5.5), (5.6), (5.8) numerically as follows We set $x_0 = 1$ and we used Eqs. (5.5) and (5.6) to determine φ and m. The set of φ, m that satisfied *at the same time* Eq. (5.8) for $x_0 = 1$, and for an (arbitrary) value of $\ln v/a^3$ ($= -0.16$ in Fig. 3), defines a *folding* line. Below this folding line $x_0 < 1$ and the system can freeze into a small number of thermodynamically dominant folds.

To proceed with analyzing the phase diagram, we note that $m = \langle\langle\sigma_i\rangle_{\rm th}\rangle_{\rm av}$ and $\varphi = \langle\langle\sigma_i\rangle_{\rm th}^2\rangle_{\rm av}$, as we explained in Sec. V A. As it is seen in Fig. 3, for $T > D\sqrt{\rho/v}$ and $T > D_0\rho$ (located above the AT line and the paramagnetic line $T = D_0\rho$) the only possible values are $\varphi = 0, m = 0$. For these values Eq. (5.8) cannot be satisfied (there is no $x_0 < 1$) and the system is a globule without a stable native structure. Since $\varphi = 0 \rightsquigarrow \langle\sigma_i\rangle_{\rm th} = 0$, the monomers interchange freely between secondary-structure states with no net preference. In Fig. 3 we designated this phase as a *disordered globule*. For $D_0 > [D\sqrt{1/(\rho v)}]$, as the temperature drops below $T = D_0\rho$ the system undergoes first a ferromagnetic transition with $m \neq 0$. From the relation $\varphi = \langle\langle\sigma_i\rangle_{\rm th}^2\rangle_{\rm av}$ it follows that $\varphi \neq 0$ as well. This ferromagnetic region extends down to the AT line. Using Eqs. (5.5), (5.6), and (5.8), one can determine a set of values φ, m that define a folding line $x_0 = 1$. The ferromagnetic phase enclosed between the paramagnetic line $T = D_0\rho$, the AT line, and the folding line corresponds to a globule with no native structure, but with a net secondary structure (mostly $+1$ or -1). This is depicted in Fig. 3 as the *ordered globule* phase. Note that the nonzero value for the rescaled variable φ merely indicates the ferromagnetic transition in the $\{\sigma\}$ variables. Strictly speaking, the rescaled variable φ is defined only for $x_0 < 1$, i.e., below the folding line. The original order parameter $\varphi_{\alpha\beta}(\mathbf{R}) = \langle\Psi_{\alpha\beta}(\mathbf{R})\rangle = 0$ $\forall\alpha,\beta$ in this region, because any replicas α,β belong to different folds. Below this phase there is a region enclosed between the folding line and the AT line. This region corresponds to an *ordered frozen* phase where the system has a stable native fold and a prevailing secondary structure.

For $D_0 < [D\sqrt{1/(\rho v)}]$ and *below* the AT line the replica symmetry inside folds is broken. The folding line cannot be calculated analytically away from the temperature $T = D\sqrt{\rho/v}$. For this reason we drew it schematically with a short-dashed line. In doing so we assumed that the folding line extends in a continuous manner below the AT line. This is justified by the fact that the transition of the φ variables along the AT line has to be continuous by analogy with the SK model. If the folding line were discontinuous, this would result to a jump in φ. The region that lies below the folding line, and is enclosed between the folding line, the AT line, and the vertical line $D_0/[D\sqrt{1/(\rho v)}] = 1$ corresponds to a phase with a native structure and a ferromagnetic secondary structure with broken symmetry [spin-glass- *(SG) ordered frozen* phase]. This means that the various monomers have a net preference for a particular state ($+1$ or -1), but this preference (unlike the ferromagnetic case) changes from monomer to monomer. If one sums over all monomers (this is equivalent to performing the average $\langle\rangle_{\rm av}$), one will find that a particular state is prevailing. Thus the system will be in a native state with both secondary structural motifs coexisting, but with one of them predominating.

In the region $T < D\sqrt{\rho/v}$ and $D_0/[D\sqrt{1/(\rho v)}] < 1$ the RSB solution gives $\varphi \neq 0$ (below the folding line) and $m = 0$. The folding line depends on φ and m. Since m stays constant ($m = 0$) in this region [27], the folding line is horizontal and equal to the value that it has at $D_0 = 0$. In this region and below the folding line the system has a native structure. From the above values of φ and m it follows that $\langle\sigma_i\rangle_{\rm th} \neq 0$ but $\langle\langle\sigma_i\rangle_{\rm th}\rangle_{\rm av} = 0$. Thus the system has an equal amount of both kinds of secondary structure (*spin-glass frozen* phase).

In the region that lies below the AT line and above the folding line the system can interchange between many folds. As we explained in Sec. V B, the existence of many different, thermodynamically dominant folds causes the variable φ to have a zero value, even though the $\{\sigma\}$ variables can freeze in a spin-glass phase for $T < D\sqrt{\rho/v}$. Our analysis in Sec. V B was carried out for the case $D_0 = 0$, but the same result holds for $D_0 \leq [D\sqrt{1/(\rho v)}]$. For $D_0 > [D\sqrt{1/(\rho v)}]$ the order parameter $m > 0$ [27], and thus $\varphi > 0$ as well. Since (in principle) the $\{\sigma\}$ variables can freeze in a spin-glass phase for $T < D\sqrt{\rho/v}$, the thermal average of $\{\sigma\}$ for every residue will depend on the local contacts (i.e., on the particular fold). Thus, as the system interconverts between folds, the secondary structure changes in a manner that depends on each fold. We designated this phase as a *spin-glass globule*.

VII. DISCUSSION AND CONCLUSIONS

In this work we investigated a model of a heteropolymer with secondary structure. We solved the model at the level of mean field theory and we determined a variety of phases. The stability of these phases depends on the temperature T, the heterogeneity of interactions expressed by the parameter D, the mean value of interactions D_0, and the polymer flexibility $\ln v/a^3$. At a high temperature the polymer is predicted to exist in a *disordered globule* state, interchanging freely between different folds and secondary-structure states. As the temperature is lowered (or equivalently the interaction parameters D, D_0 become larger) the polymer undergoes first a phase transition that leads to stable secondary structure, and then a transition to a *frozen fold* phase with a few [of order $O(1)$] thermodynamically dominant folds. The secondary-structure transition can be ferromagnetic or spin-glass-like, depending on the relative strength of the parameters D and D_0. In the former case there is a single preferred secondary-structure state throughout the polymer, whereas in the second case the dependence for a secondary-structure state depends on the monomer, with both states present. The transition to the frozen fold phase is governed by the flexibility parameter $\ln v/a^3$.

The heteropolymeric nature of the model was represented by a set of quenched disorder variables D_{ij} that obeyed a Gaussian distribution. This representation, referred to as the *independent interaction model* [11], implies that the interaction between a pair of residues depends on the particular pair. In an alternative representation, one can model the polymer sequence by a set of quenched random variables σ_i. In that case, the interaction strength between residues i and j in contact will be $\sigma_i\sigma_j$. In this representation the interaction depends on the individual character of the residues, and the interaction strengths $\sigma_i\sigma_j$ are correlated. This has been referred to previously as the *sequence model* and has been solved for the case of the two-letter code [11].

The secondary structure was modeled by a set of internal states that the monomers occupied. As we discussed in Sec. II, this representation does not relate explicitly the secondary structure to the detailed geometry of a fold. It is possible to attribute a geometric nature to this model of secondary structure, if one considers that a fold is defined up to the *characteristic scale* v associated with the monomer specific volume. The energy of a fold depends then on (a) the contacts made between residues up to this scale, and (b) on some additional characteristics of the residues in contact (their internal states), that might have to do with their relative orientation, backbone dihedral angles, etc. An important feature of the above model is that formation of a *stable* secondary structure is energetically favored and entropically hindered. Thus this model captures an essential property of the formation of secondary structure observed in nature.

The interactions between residues were considered to be short range in space, but long range in sequence (that is, any two residues on the sequence could interact, provided they were in contact). This means that in our model the interaction between two residues is affected by their secondary-structure state no matter how far away they are in sequence. One could argue that this assumption is more realistic for the case of β-sheet formation, whereas α-helical segments are stabilized by interactions between neighboring in sequence residues. It is a well established fact though [12,34,35] that helices are significantly stabilized by interactions of nonpolar residues that are distant in sequence and close in space. Due to this reason we think that it is not necessary to treat differently the near neighbor (in sequence) from the long-range interactions, when trying to reproduce some basic features of secondary structure.

The stabilization of secondary-structure elements depends on the mean interaction D_0. For a positive value of D_0 a residue in a certain state will interact favorably with contact residues in the same state. On the other hand, if $D_0 = 0$ a residue will stabilize on the average an equal number of contact residues with the same and with opposite secondary structure. It is possible that the variability in terms of secondary-structure motifs observed in different proteins [36] (i.e., all α helical or all β sheet, as opposed to mixed α-β) is caused in a similar way through stabilization of a particular secondary-structure motif by (respectively) the same or the opposite motif.

As it follows from the solution, in this model the polymer is able to freeze into particular folds after the residues adopt stable secondary-structure states. Each fold is associated with a *free* energy level, because it corresponds to a set of conformations of the $\{\sigma_i\}$ variables. Since the folds (in the low free energy part of the spectrum) are completely different of each other, these free energy levels are independent of each other. This is similar to what was observed in previous works for the random heteropolymer [7,8] and the random energy model [10]. The overlap of $\{\sigma_i\}$ variables depends on the temperature. This means that the free energy levels will change with temperature. Thus in this model it is possible that the ground state configuration (as a function of both the spatial coordinates $\{\mathbf{r}_i\}$ and the $\{\sigma_i\}$ variables) changes with temperature.

It is interesting to compare this model with other models of protein folding, which encapsulate the concept of *minimal frustration* [5,6]. This concept states that the secondary-structure propensities are not in conflict with the native tertiary structure. In our model, specific folds become thermodynamically important when the mean value of the interaction strength $D_0 \geq 0$. This means that the residue contacts (i.e., the native fold) affect the secondary-structure states, and the energy is lowered on the average when the residues adopt the correct secondary-structure states (for a particular fold). Thus our model is in accordance with minimal frustration. However, in our case the *individual* interaction strengths D_{ij} obey a probability distribution that is nonzero for $D_{ij} < 0$. Thus a residue will have both positive and negative interaction strengths with its neighboring residues, leading inevitably to some frustrated secondary-structure states. This is equivalent to the frustration observed in the spin-glass systems.

A different model of secondary-structure formation has been introduced in [33]. In this work the polymer is embedded on a hypercubic lattice. The geometric characteristics of the secondary structure are taken into account by modeling the α-helical state as a straight line and assigning an energetic penalty to turn formation, that breaks the helix. The model predicts a low temperature transition to a "frozen" phase in which the helix extends throughout the entire sequence, with the exception of forming turns on the surface. In other words, in this phase the polymer consists of fully stretched paths that turn on the surface of the lattice (to retain compactness). A disadvantadge of the model is that it refers to a homopolymer. Thus this low temperature "frozen" phase does not really correspond to a unique tertiary structure, because the helical segments can be rearranged in many ways, retaining the same compactness and corresponding to the same energy.

One important question that has to be addressed is the validity of the mean field theory approach followed in solving this problem. Our model is in reality a short-range spin glass, with the additional property that due to its polymeric nature, the "spins" (i.e., the residues) are allowed to sample over many different neighboring contacts. From this point of view each "spin" is able to feel the influence of many other "spins," behaving in a similar manner as in a long-range spin glass.

The freezing pattern in the space of folds is derived by a mean field approximation for the variables $Q_{\alpha\beta}$. This treatment has been shown to give satisfactory results in the case of simple heteropolymers [7,8], leading to a REM picture for the energy spectrum. This picture has been confirmed in exact enumeration studies of lattice models with a simple heteropolymer Hamiltonian [37]. The contribution of fluctuations in the order parameter $Q_{\alpha\beta}$ has been examined for the two-letter code in [11]. It has been shown that the effect of fluctuations is to reduce the freezing into a particular fold temperature.

The mean field assumption enabled us to relate the problem with the SK free energy, with order parameters given by $m_\alpha = \sum_i \sigma_i^\alpha/N$ and $\varphi_{\alpha\beta} = \sum_i \sigma_i^\alpha \sigma_i^\beta/N$. The free energy was found to depend in addition on the order parameter x_0 that determines the degree of freezing to a particular fold. The solution is always replica-symmetry breaking (below a certain temperature) in terms of the variables $\varphi_{\alpha\beta}$, because $\varphi_{\alpha\beta} = 0$ always for α, β belonging to *different* folding groups. Being consistent with the analogy with the SK model, we chose the RSB solution for $\varphi_{\alpha\beta}$ *inside each folding group* below the AT line, and we determined the $\varphi_{\alpha\beta}$ pattern using the Parisi ansatz. According to this pattern, the $\varphi_{\alpha\beta}$ overlap varies inside each group from a minimum (0) to a maximum value that depends on temperature, approaching the limiting value 1 as $T \to 0$ [32]. This means that there are always many thermodynamically important pure states in terms of the σ_i variables.

A considerable amount of work has been oriented towards elucidating the properties of short-range spin-glass systems. More specifically, it has been argued with the help of scaling arguments that short-range spin glasses have *exactly one* pair of ground states [38]. The existence of an AT transition for any finite dimensionality spin glasses has also been questioned [38,39]. Short-range spin-glass systems have also been studied with momentum-space [9,40] and real-space [9,41] renormalization-group methods and Monte Carlo calculations [9]. The upper critical dimension, above which mean field is exact, has been shown [40] to be $d_u = 6$. On the other hand, real-space renormalization-group (RG) studies [41] have shown that the phase diagram for a short-range spin glass with $\pm J$ interactions on a cubic lattice (in $d = 3$) is similar to the predictions of mean field theory, with paramagnetic, ferromagnetic, antiferromagnetic (for $+J$ predominating), and spin-glass phases. The spin-glass transition temperature for a cubic lattice ($d = 3$) was found [9,42] to be $T_f \simeq T_f^{\rm MF}/2$, with $T_f^{\rm MF} = z^{1/2}\Delta J$, z the coordination number, and ΔJ the width of the interaction. Also, the case of *dilute* infinite-ranged spin glasses has been investigated in [43]. In this work the spin-spin interactions were strong (of order unity, as in our model), but each spin interacted with a fraction p/N of other spins, with p finite. The resulting phase diagram had paramagnetic, ferromagnetic, spin-glass, and mixed phases.

If the above picture is representative of a heteropolymer with secondary structure, to which region (if any) of the phase diagram of Fig. 3 do proteins belong? To answer this question one should keep in mind that proteins are much more complex systems than what this model implies. For example, the distinction between secondary-structure "states" is probably not as sharp as an Ising-like representation dictates. The solvent conditions, or intrinsic preferences of the various residues for a particular secondary-structure motif, might destroy the $+1 \to -1$ degeneracy of the $\{\sigma\}$ variables satisfied in our model. This would necessitate the introduction of a field in our Hamiltonian. If we consider the above model as a first approximation to a protein description, we can argue that various proteins belong to different regions in the phase diagram. In nature there are proteins which are α helical or β sheet, or have both secondary structural motifs. The first two kinds would be represented by the right part of the phase diagram, where the secondary-structure formation is described by a ferromagnetic transition. The third kind would correspond to the spin-glass region.

ACKNOWLEDGMENTS

We acknowledge helpful discussions with Alexander Gutin and Chryssostomos Sfatos. This work was supported partially by the David and Lucille Packard Foundation. G. A. wishes to acknowledge Martin Karplus for support.

APPENDIX

In this appendix we show that higher-order terms in the expansion of Eq. (4.1) do not affect the pattern of freezing in the $\{r\}$ space. To show this, it suffices to demonstrate that these terms scale as powers $1/R^\nu$ of the characteristic scale R, with $\nu \geq 3$. In this case the entropy term ($\sim 1/R^2$) becomes more significant in the regime $R \to \infty$ and the free energy $\mathcal{F}$ has a barrier between $R = \infty$ (no freezing) and $R = v^{1/3}$ (freezing in microscopic scales).

To examine the effect of higher-order terms we have to perform the integration over the $\{\varphi_{\alpha\beta}\}$ variables. The only surviving contributions come from terms that include even powers of $\{\varphi_{\alpha\beta}\}$. Equation (4.1) becomes

$$\langle Z(\{D_{ij}\})^n \rangle_{\rm av} = \int \mathcal{D}m_\alpha(\mathbf{R}) \exp\left\{ -\frac{\beta D_0}{2} \sum_\alpha \int d\mathbf{R} m_\alpha^2(\mathbf{R}) \right\}$$

$$\times \int \prod_{i,\alpha} d\mathbf{r}_i^\alpha g(\mathbf{r}_{i+1}^\alpha - \mathbf{r}_i^\alpha) \exp\left\{ -\frac{1}{2} \sum_{\alpha<\beta} \int d\mathbf{R}_1 d\mathbf{R}_2 \ln[1 - (\beta D)^2 Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2)] \right\}$$

$$\times\Bigg\{1+\frac{(\beta D)^8}{48}\sum_{\{\alpha,\beta\}}\Big\langle\int\varphi^4_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)\Big\rangle$$
$$+\frac{(\beta D)^8}{32}\sum_{\{\alpha,\beta,\gamma,\delta\}}\Big\langle\int\varphi^2_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)\int\varphi^2_{\gamma\delta}(\mathbf{R}_3,\mathbf{R}_4)Q_{\gamma\delta}(\mathbf{R}_3,\mathbf{R}_4)\Big\rangle$$
$$+\frac{(\beta D)^8}{32}\sum_{\{\alpha,\beta,\gamma\}}\Big\langle\int\varphi^2_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)\varphi^2_{\beta\gamma}(\mathbf{R}_2,\mathbf{R}_3)Q_{\alpha\beta\gamma}(\mathbf{R}_1,\mathbf{R}_2,\mathbf{R}_3)\Big\rangle$$
$$+\frac{(\beta D)^4(\beta D_0)^2}{8}\rho\sum_{\epsilon}\int m^2_{\epsilon}(\mathbf{R})\sum_{\{\alpha,\beta\}}\Big\langle\int\varphi^2_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)\Big\rangle+\cdots\Bigg\}, \tag{A1}$$

where we omitted the terms that did not contain the variables $Q_{\alpha\beta\gamma\ldots}$. In Eq. (A1) we denoted with $\langle\,\rangle$ the averages

$$\langle A[\varphi_{\alpha\beta}]\rangle=\frac{\int d\varphi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)A[\varphi_{\alpha\beta}]e^{-\Lambda[\varphi_{\alpha\beta}]}}{\int d\varphi_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)e^{-\Lambda[\varphi_{\alpha\beta}]}},$$

with $\Lambda[\varphi_{\alpha\beta}]$ defined as

$$\Lambda[\varphi_{\alpha\beta}]=\frac{(\beta D)^2}{2}\sum_{\alpha<\beta}\int d\mathbf{R}_1d\mathbf{R}_2\big[1-(\beta D)^2Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)\big]\varphi^2_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2).$$

After evaluating these averages we get

$$\langle Z(\{D_{ij}\})^n\rangle_{\mathrm{av}}=\int\mathcal{D}m_\alpha(\mathbf{R})\exp\Big\{-\frac{\beta D_0}{2}\sum_\alpha\int d\mathbf{R}m^2_\alpha(\mathbf{R})\Big\}\int\prod_{i,\alpha}d\mathbf{r}^\alpha_i g(\mathbf{r}^\alpha_{i+1}-\mathbf{r}^\alpha_i)$$
$$\times\exp\Big\{-\frac{1}{2}\sum_{\alpha<\beta}\int d\mathbf{R}_1d\mathbf{R}_2\ln[1-(\beta D)^2Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)]\Big\}$$
$$\times\Bigg\{1+\frac{3}{4}\frac{(\beta D)^8}{48}\sum_{\alpha\beta}\int\frac{Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)}{\big[1-(\beta D)^2Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)\big]^2}$$
$$+\frac{1}{4}\frac{(\beta D)^8}{32}\sum_{\{\alpha,\beta,\gamma,\delta\}}\int\frac{Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)}{1-(\beta D)^2Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)}\int\frac{Q_{\gamma\delta}(\mathbf{R}_3,\mathbf{R}_4)}{1-(\beta D)^2Q_{\gamma\delta}(\mathbf{R}_3,\mathbf{R}_4)}$$
$$+\frac{1}{4}\frac{(\beta D)^8}{32}\sum_{\{\alpha,\beta,\gamma\}}\int\frac{Q_{\alpha\beta\gamma}(\mathbf{R}_1,\mathbf{R}_2,\mathbf{R}_3)}{\big[1-(\beta D)^2Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)\big]\big[1-(\beta D)^2Q_{\alpha\gamma}(\mathbf{R}_1,\mathbf{R}_3)\big]}$$
$$+\frac{1}{2}\frac{(\beta D)^4(\beta D_0)^2}{8}\rho\sum_{\epsilon}\int m^2_{\epsilon}(\mathbf{R})\sum_{\{\alpha,\beta\}}\int\frac{Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)}{1-(\beta D)^2Q_{\alpha\beta}(\mathbf{R}_1,\mathbf{R}_2)}+\cdots\Bigg\}. \tag{A2}$$

Upon reexponentiating and expanding the denominators of the various terms, and using the scaling relation (4.3), it is directly verified that the resulting contributions to the free energy scale as ≥ 3 powers of $1/R$.

[1] *Protein Folding*, edited by T. Creighton (Wiley, New York, 1992).
[2] C. Levinthal, J. Chim. Phys. **65**, 44 (1968).
[3] D. Bashford, D. L. Weaver, and M. Karplus, J. Biomol. Struct. Dyn. **1**, 1243 (1984).
[4] M. Karplus and D. Weaver, Biopolymers **18**, 1421 (1979).
[5] J. B. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).
[6] J. B. Bryngelson and P. G. Wolynes, Biopolymers **30**, 177 (1990).
[7] A. M. Gutin and E. I. Shakhnovich, Zh. Eksp. Teor. Fiz. **96**, 2096 (1989) [Sov. Phys. JETP **69**, 1185 (1989)].
[8] E. I. Shakhnovich and A. M. Gutin, Biophys. Chem. **34**, 187 (1989).
[9] K. Binder and A. P. Young, Rev. Mod. Phys. **58**, 801 (1986).
[10] B. Derrida, Phys. Rev. Lett. **45**, 79 (1980).
[11] C. D. Sfatos, A. M. Gutin, and E. I. Shakhnovich, Phys. Rev. E **48**, 465 (1993).
[12] *Proteins*, edited by T. E. Creighton (W. H. Freeman and Co., New York, 1993), p. 255.
[13] J. M. Thornton, in [1], Chap. 2.
[14] H. S. Chan and K. A. Dill, J. Chem. Phys. **90**, 492 (1990).
[15] P. Y. Chou and G. D. Fasman, Annu. Rev. Biochem. **47**, 251 (1978).
[16] O. B. Ptitsyn and A. V. Finkelstein, Biopolymers **22**, 15

(1983).
[17] G. E. Schulz, Annu. Rev. Biophys. Biophys. Chem. **17**, 1 (1988).
[18] K. F. Freed, *Renormalization Group Theory of Macromolecules* (Wiley, 1987, New York, 1987), Chap. 5.
[19] I. M. Lifshitz, A. Yu. Grosberg, and A. R. Khokhlov, Rev. Mod. Phys. **50**, 683 (1978).
[20] E. I. Shakhnovich and A. M. Gutin, J. Phys. A **22**, 1647 (1989).
[21] M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[22] P. G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, NY, 1979).
[23] S. Kirkpatrick and D. Sherrington, Phys. Rev. B **17**, 4384 (1978).
[24] J. R. L. de Almeida and D. J. Thouless J. Phys. A **11**, 129 (1978).
[25] G. Parisi, J. Phys. A **13**, 1887 (1980).
[26] G. Parisi, J. Phys. A **13**, 1101 (1980).
[27] G. Toulouse, J. Phys. (Paris) Lett. **41**, L447 (1980).
[28] A. J. Bray and M. A. Moore, Phys. Rev. Lett. **41**, 1068 (1978).
[29] E. Pytte and J. Rudnick, Phys. Rev. B **19**, 3603 (1979).
[30] D. J. Thouless, J. R. de Almeida, and J. M. Kosterlitz, J. Phys. C **13**, 3271 (1980).
[31] G. Parisi, J. Phys. A **13**, 1101 (1980).
[32] J. Vannimenus, G. Toulouse, and G. Parisi, J. Phys. (Paris) **42**, 565 (1981).
[33] J. Bascle, T. Garel, and H. Orland, J. Phys. (Paris) **3**, 259 (1993).
[34] W. F. DeGrado, Z. R. Wasserman, and J. D. Lear Science **243**, 622 (1989).
[35] A. Rey and J. Skolnick, *Proteins* **16**, 8 (1993).
[36] J. S. Richardson, Adv. Prot. Chem. **34**, 167 (1981).
[37] E. I. Shakhnovich and A. M. Gutin, J. Chem. Phys. **93**, 5967 (1990).
[38] D. S. Fisher and D. A. Huse, J. Phys. A **20**, L1005 (1987).
[39] D. S. Fisher and D. A. Huse, Phys. Rev. B **38**, 386 (1988).
[40] A. B. Harris, T. C. Lubensky, and J. H. Chen, Phys. Rev. Lett. **36**, 415 (1976).
[41] C. Jayaprakash, J. Chalupa, and M. Wortis, Phys. Rev. B **15**, 1495 (1977).
[42] B. W. Southern and A. P. Young, J. Phys. C **10**, 2179 (1977).
[43] L. Viana and A. J. Bray, J. Phys. C **18**, 3037 (1985).