Proteins with selected sequences: A heteropolymeric study

J. Wilder* and E. I. Shakhnovich†
Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138

(Received 5 May 2000)

Protein sequences are expected not to be random but selected in order to form a stable native structure that is kinetically accessible. Therefore our model contains a selective temperature in sequence space (see [S. Ramanathan and E. Shakhnovich, Phys. Rev. E **50**, 1303 (1994)]) to optimize the sequence for the target conformation statistically. Replica calculations, which go beyond quadratic approximations in the field-theoretical Hamiltonian, are presented. A phase diagram indicating the temperatures and selective temperatures at which transitions to a frozen globule, i.e., the native state, occur is obtained. It is shown that going beyond the quadratic approximation in the field Hamiltonian is very important, since it results in a significant change of the phase diagram. Moreover, we suggest that a one-step replica permutation symmetry scheme is sufficient to solve the model. In addition to this we present a result for the sequence correlation function along the chain in the case of a short-ranged potential between the monomers. A correlation function between monomers that form a contact in the native state is given depending on the temperature and the interaction parameter.

PACS number(s): 87.15.Nn, 64.60.Cn, 64.60.Kw

I. INTRODUCTION

Real proteins are composed out of a set of 20 different kinds of amino acids, which leads to a complex interaction potential between the monomers of the protein. Since it is impossible to deal with all these interactions analytically, the statistical-mechanical approach to protein folding is based on the investigation of simple models of heteropolymers. A first step in this context is an approach where the heteropolymers are made up of only two kinds of monomers; thus similar monomers attract each other, whereas unlike monomers repel each other. These two monomers represent hydrophilic (A) and hydrophobic (B) amino acids. This simplified model for proteins still supports the important feature that the interaction energy is characterized by the sequence of the monomers of the protein, which is an advantage over the independent interaction model, where the interaction energies are considered as independent [2–6].

The statistical mechanics of random copolymers consisting of two different kinds of monomers has been studied in previous papers [6–11]. Though different kinds of monomers repel each other, the monomers are not able to separate into hydrophobic- and hydrophilic-rich macroscopic regions at low temperatures because of the presence of chain connectivity. The positions of neighbors are not independent, which leads to frustration.

It is suggested in [10] using the one-step replica symmetry breaking ansatz that the energy levels show a continuous spectrum for large values of the energy and a discrete spectrum for lower values. As in the independent interaction model the system freezes into the lower part of the energy spectrum if the temperature is decreased sufficiently. The ability, however, to fold into a kinetically accessible unique native state requires that the energy level of this state be far below the discrete part of the energy spectrum of the chain [14]. Such an energy spectrum can be designed by the introduction of phenomenological models, which are motivated by the idea of nonrandomness in proteins [2,12,13]. These models favor native-state contacts energetically, thus pulling down the energy of the native structure.

There is, however, the caveat that the fundamental interaction in proteins is basically the same as in random heteropolymers. According to [14] the ability of folding into unique structure might be achieved by the optimization of the sequence of amino acids. In [1] the statistical-mechanical properties of heteropolymers with designed sequences are investigated. The sequences are designed in the sense that a canonical distribution in sequence space depending on the native conformation and a selective temperature T_s weights the probability of the occurrence of a certain sequence. Consequently some sequences are favored over others.

A phase diagram within Gaussian approximation for such a heteropolymer depending on the selective temperature T_s and the real temperature T is calculated in [1] using a one-step replica symmetry breaking ansatz. In the present work we show that this phase diagram is changed significantly by going beyond the Gaussian approximation after investigating the question as to whether a higher-order replica symmetry breaking ansatz is necessary. The calculations show that a symmetry breaking ansatz of higher order reduces to a one-step replica symmetry breaking scheme.

Finally we present a sequence correlation function for a short-ranged potential between the monomers. The dependence on the strength of the interaction and the polymerization temperature is analyzed.

II. MODEL AND DEFINITIONS

Let the set $\{\mathbf{r}_i\}$ represent the conformation of the heteropolymeric chain, where the index i refers to the ith monomer. Then the interaction term in the Hamiltonian reads

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

*Email address: juergen@chasma.harvard.edu
†Email address: eugene@diamond.harvard.edu

with N the number of monomers in the chain and $U(\mathbf{r}_i - \mathbf{r}_j)$ a short-ranged potential. The binary interaction virial coefficients are given by [15]

$$b_{ij} = 2[b_0 + A(\sigma_i + \sigma_j) + \chi\sigma_i\sigma_j], \qquad (2)$$

where the sequence of monomers is described by the set of variables $\{\sigma_i\}$. If monomer i is of type A, then $\sigma_i = 1$ and $\sigma_{\rm i} = -1$ in case that this monomer is of type B. Here we are interested in the case that similar monomers attract each other, which means that the Flory parameter χ is smaller than zero. The parameter A vanishes if the interactions between similar monomers are equal, which is assumed in the following. To make sure that we are dealing with compact globular states, b_0 is set to negative values, which implies a certain overall attraction independent of the specific sequence of the monomers.

The set $\{\mathbf{r}_i^0\}$ represents the conformation of the target or native structure for which the sequence of monomers should be designed. The potential energy of the chain consisting of the sequence $\{\sigma_i\}$ folded to the native structure is given by

$$\mathcal{H}_0(\{\sigma_i\}) = \frac{1}{2}\sum_{i,j}^{N} b_{ij} U(\mathbf{r}_i^0 - \mathbf{r}_j^0). \qquad (3)$$

In the canonical ensemble of sequences in sequence space according to this Hamiltonian we get the following distribution function for sequence sets $\{\sigma_i\}$ [14]:

$$P\{\sigma_i\} = \frac{1}{\tilde{Z}}\exp\left(-\frac{\mathcal{H}_0(\{\sigma_i\})}{T_s}\right), \qquad (4)$$

where the sequence space partition function is given by

$$\tilde{Z} = \sum_{\{\sigma_i\}} \exp\left(-\frac{\mathcal{H}_0(\{\sigma_i\})}{T_s}\right), \qquad (5)$$

with the polymerization temperature or selective temperature T_s. The sequence is treated as a frozen disorder; thus to get the free energy we have to average the logarithm of the partition function in the conformational space over the probability distribution given by Eq. (4). So the free energy reads

$$F = -kT\langle \ln Z\{\sigma_i\}\rangle_{\rm av} = -kT\sum_{\{\sigma_i\}} \ln Z\{\sigma_i\} P\{\sigma_i\}. \qquad (6)$$

$\langle \cdots \rangle_{\rm av}$ denotes averaging over all possible sequences $\{\sigma_i\}$ with a probability distribution $P\{\sigma_i\}$. The conformational partition function Z for a given sequence $\{\sigma_i\}$ can be written as

$$Z = \sum_{\{\mathbf{r}_i\}}\left[\exp\left(-\frac{\mathcal{H}(\{\mathbf{r}_i\})}{kT}\right)\right]\prod_i g(\mathbf{r}_{i+1} - \mathbf{r}_i), \qquad (7)$$

where the function g for next neighbors along the chain, which ensures the connectivity of the chain in its standard Gaussian form, is given by [16]

$$g(\mathbf{r}_{j+1} - \mathbf{r}_j) = \frac{1}{(2\pi a^2)(3/2)}\exp\left[-\frac{(\mathbf{r}_{j+1} - \mathbf{r}_j)^2}{2a^2}\right], \qquad (8)$$

with a the segment length in the discrete polymer model.

The calculation of the free energy according to Eq. (6) is very difficult, caused by the occurrence of the logarithm in this equation. To solve this problem we make use of the replica trick (see, e.g., [17]), which requires the calculation of the averaged nth power of the partition function

$$\langle Z^n\rangle_{\rm av} = \sum_{\{\sigma_i\}} \int \mathcal{D}\mathbf{r}_j^\alpha g(\mathbf{r}_{j+1}^\alpha - \mathbf{r}_j^\alpha)\exp\left(-\frac{b_0}{T}\sum_{i,j} U_{ij}^\alpha\right)$$
$$\times \exp\left[-\sum_{\alpha=1}^{n}\sum_{i,j}\frac{\chi}{T}\sigma_i U(\mathbf{r}_i^\alpha - \mathbf{r}_j^\alpha)\sigma_j\right] P\{\sigma_i\}, \qquad (9)$$

with $U_{ij}^\alpha = U(\mathbf{r}_i^\alpha - \mathbf{r}_j^\alpha)$ and $\mathbf{r}_i^\alpha$ the position of the ith monomer in replica α. Since U is a nonlocal potential, we perform a Hubbard-Stratonovich transformation and calculate the trace over σ_i. By setting

$$\phi_\alpha(\mathbf{R}_2) = \frac{1}{b_\alpha}\int d\mathbf{R}_1 U^{-1}(\mathbf{R}_1 - \mathbf{R}_2)\psi_\alpha(\mathbf{R}_1), \qquad (10)$$

where ψ_α is the Hubbard-Stratonovich field in the αth replica, b_α equals $-\chi/T_s$ for the zeroth replica ($\alpha = 0$), and $-\chi/T$ for all other replicas, we get

$$\langle Z^n\rangle_{\rm av} = \frac{1}{\tilde{Z}}\Bigg\langle \int \mathcal{D}\Phi_\alpha \mathcal{D}\Phi_0 \exp\Bigg[-\sum_{\alpha=0}^{n} b_\alpha \int d\mathbf{R}_1 d\mathbf{R}_2 \Phi_\alpha(\mathbf{R}_1)\Phi_\alpha(\mathbf{R}_2)$$
$$\times U(\mathbf{R}_1 - \mathbf{R}_2) + 2\sum_{\alpha,\beta=0}^{n} b_\alpha b_\beta \int d\mathbf{R}_1 d\mathbf{R}_1' \Phi_\alpha(\mathbf{R}_1') U(\mathbf{R}_1 - \mathbf{R}_1')$$
$$\times \int d\mathbf{R}_2 d\mathbf{R}_2' \Phi_\alpha(\mathbf{R}_2') U(\mathbf{R}_2 - \mathbf{R}_2') Q_{\alpha\beta}(\mathbf{R}_1 - \mathbf{R}_2)$$
$$-\frac{4}{3}\sum_{\alpha,\beta,\gamma,\delta} b_\alpha b_\beta b_\gamma b_\delta \int d\mathbf{R}_1 d\mathbf{R}_1' \cdots d\mathbf{R}_4 d\mathbf{R}_4' \Phi_\alpha(\mathbf{R}_1') U(\mathbf{R}_1 - \mathbf{R}_1') \cdots \Phi_\delta(\mathbf{R}_4') U(\mathbf{R}_4 - \mathbf{R}_4')$$
$$\times \sum_i \delta(\mathbf{r}_i^\alpha - \mathbf{R}_1)\delta(\mathbf{r}_i^\beta - \mathbf{R}_2)\delta(\mathbf{r}_i^\gamma - \mathbf{R}_3)\delta(\mathbf{r}_i^\delta - \mathbf{R}_4)\Bigg]\Bigg\rangle_{\rm th}. \qquad (11)$$

The thermal average $\langle \cdots \rangle_{\mathrm{th}}$ includes the integrals over $\{\mathbf{r}_i^0\}$ and $\{\mathbf{r}_i^\alpha\}$, which are not given explicitly in Eq. (11). The parameter $Q_{\alpha\beta}$ describes the overlap between replicas and is defined as

$$Q_{\alpha\beta}(\mathbf{R}_1-\mathbf{R}_2)=\sum_i \delta(\mathbf{r}_i^\alpha-\mathbf{R}_1)\,\delta(\mathbf{r}_i^\beta-\mathbf{R}_2). \tag{12}$$

Neglecting terms of order χ^4, which means considering the case of weak interactions and assuming a δ function for the short-ranged potential U,

$$U(\mathbf{R}_1-\mathbf{R}_2)=\delta(\mathbf{R}_1-\mathbf{R}_2), \tag{13}$$

Eq. (11) can be evaluated by switching to $Q_{\alpha\beta}$ variables, which yields

$$\langle Z^n\rangle_{\mathrm{av}}=\int \mathcal{D}Q_{\alpha\beta}\exp[-E\{Q_{\alpha\beta}\}+S\{Q_{\alpha\beta}\}], \tag{14}$$

where the effective energy term E in the Gaussian approximation is given by

$$E\{Q_{\alpha\beta}\}=\ln\int \mathcal{D}\phi_\alpha(\mathbf{k})\exp\left[-V\sum_{\alpha,\beta=0}^{n}\sum_{\mathbf{k}}[b_\alpha\delta_{\alpha\beta}\right.$$
$$\left.-2b_\alpha b_\beta Q_{\alpha\beta}(\mathbf{k})]\phi_\alpha(\mathbf{k})\phi_\beta(-\mathbf{k})\right], \tag{15}$$

with V the volume of the system and $\mathbf{k}$ is the wave vector. The entropy S, which corresponds to the change of the variables, reads

$$S\{Q_{\alpha\beta}\}=\ln\left\langle \delta\left(Q_{\alpha\beta}(\mathbf{R}_1-\mathbf{R}_2)-\sum_i \delta(\mathbf{r}_i^\alpha-\mathbf{R}_1)\right.\right.$$
$$\left.\left.\times\,\delta(\mathbf{r}_i^\beta-\mathbf{R}_2)\right)\right\rangle_{\mathrm{th}}. \tag{16}$$

For more details of the derivation of Eqs. (11) and (15) see [1]. Performing the integrals over the field $\phi_\alpha(\mathbf{k})$ for $\alpha=0,\ldots,n$ in Eq. (15) one obtains

$$\int d\mathbf{k}\,\ln[\det P_{\alpha\beta}(\mathbf{k})], \tag{17}$$

with $P_{\alpha\beta}(\mathbf{k})=b_\alpha\delta_{\alpha\beta}-2b_\alpha b_\beta Q_{\alpha\beta}(\mathbf{k})$ a symmetric matrix.

III. REPLICA SYMMETRY BREAKING INVESTIGATIONS

In [1] the energy represented by Eq. (15) is calculated by a one-step replica symmetry breaking scheme. The Parisi-type hierachical matrix for the order parameter $Q_{\alpha\beta}$ is formed by dividing all the replicas in groups by the following scheme: Two replicas α and β are in the same group if they overlap on a microscopic scale, which implies that the overlap parameter $Q_{\alpha\beta}(\mathbf{R}_1-\mathbf{R}_2)=\rho\,\delta(\mathbf{R}_1-\mathbf{R}_2)$ with ρ the density of the system. The replicas α and β are in different groups if they do not overlap, which m eans that the overlap parameter vanishes, i.e., $Q_{\alpha\beta}=0$.

This one-step replica symmetry breaking scheme implies that the overlap between the replicas and the target conformation does not vary continuously but can only assume two possible values. As a consequence there is a sharp change from the disordered globule or frozen globule to the target conformation, which depends on the selective temperature and the real temperature. If a conformation in replica α folded into the target state, then we get $Q_{0\alpha}=\rho\,\delta(\mathbf{R}_1-\mathbf{R}_2)$ and $Q_{0\alpha}=0$ otherwise.

We are now going to show that this one-step replica symmetry breaking scheme, which was introduced in [1] quite intuitively, is the appropriate one for our model. It can be seen that a two-step replica symmetry breaking ansatz reduces to the one-step symmetry breaking scheme. This implies that there is indeed a sharp change from the disordered or frozen globule to the target conformation.

The energy depending on the order parameter $Q_{\alpha\beta}$ according to Eq. (15) within the replica formalism is given by [17]

$$E\{Q_{\alpha\beta}\}=\lim_{n\to 0}\frac{1}{n}\mathrm{Tr}\,\ln[P_{\alpha\beta}]. \tag{18}$$

In the disordered or frozen state, when there is no overlap with the native state, this expression can be calculated in terms of the Parisi function $a(x)$ $(x=0,\ldots,1)$ as follows [18]:

$$\lim_{n\to 0}\frac{1}{n}\,\mathrm{Tr}\,\ln[P_{\alpha\beta}]$$
$$=\ln[b_s-2b_s^2\rho]+\ln\left[b-2b^2\rho-\int_0^1 dx\,a(x)\right]-\int_0^1\frac{dx}{x^2}$$
$$\times\ln\left[\frac{b-2b^2\rho-\int_0^1 dx\,a(x)+\int_0^x dy\,a(y)-xa(x)}{b-2b^2\rho-\int_0^1 dx\,a(x)}\right]. \tag{19}$$

The entropy term S is calculated in [3] within a Gaussian approximation in terms of the Parisi function as

$$S = Nn\frac{3}{4}a^2\int_0^1 a(x)dx. \tag{20}$$

Equation (19) and (20) lead to the following mean-field free energy density, which has to be maximized [17] with respect to the Parisi function:

$$\frac{f}{n} = \ln[b_s - 2b_s^2\rho] + \ln\left[b - 2b^2\rho - \int_0^1 dx\, a(x)\right] - \int_0^1 \frac{dx}{x^2}$$

$$\times \ln\left[\frac{b - 2b^2\rho - \int_0^1 dx\, a(x) + \int_0^x dy\, a(y) - xa(x)}{b - 2b^2\rho - \int_0^1 dx\, a(x)}\right]$$

$$-\frac{3}{4}a^2\int_0^1 a(x)dx. \tag{21}$$

Now we are introducing in addition to the one step in the Parisi function, which was already done in [1] and [10], a second step according to the scheme

$$a(x) = 0 \quad \text{for} \quad x < x_0,$$

$$a(x) = a_1 \quad \text{for} \quad x_0 \leqslant x \leqslant x_1,$$

$$a(x) = a_2 = -2b^2\rho \quad \text{for} \quad x_1 < x \leqslant 1. \tag{22}$$

In contrast to the one-step replica symmetry breaking scheme we get two additional parameters a_1 and x_1. After insertion of the step function given in Eqs. (22) for the Parisi function the free energy density reads

$$\frac{f}{n} = \ln[b_s - 2b_s^2\rho] + \ln[b - (x_1 - x_0)a_1 - x_1 2b^2\rho]$$

$$+\left[\frac{1}{x_1} - \frac{1}{x_0}\right]\ln\left[\frac{b - x_1(a_1 + 2b^2\rho)}{b - (x_1 - x_0)a_1 - x_1 2b^2\rho}\right]$$

$$+\left[1 - \frac{1}{x_1}\right]\ln\left[\frac{b}{b - (x_1 - x_0)a_1 - x_1 2b^2\rho}\right]$$

$$-\frac{3}{4}a^2[(x_1 - x_0)a_1 - (1 - x_1)2b^2\rho]. \tag{23}$$

Maximizing the free energy density with respect to the parameters a_1 and x_1 leads to two coupled mean-field equations, which are calculated by setting the partial derivatives with respect to a_1 and x_1 to zero. These equations are given by

$$-\frac{1}{x_1^2}\ln[1 - x_1(a_1' + 2b\rho)]$$

$$+\frac{x_0 a_1'(a_1' + 2b\rho)}{[1 - x_1(a_1' + 2b\rho)][1 + x_0 a_1' - x_1(a_1' + 2b\rho)]}$$

$$\times\left[\frac{1}{x_1} - \frac{1}{x_0}\right] - \frac{1}{x_1}\,\frac{a_1' + 2b\rho}{1 + x_0 a_1' - x_1(a_1' + 2b\rho)}$$

$$-\frac{3}{4}a^2 b(a_1' + 2b\rho) = 0 \tag{24}$$

and

$$\left[\frac{1}{x_0} - \frac{1}{x_1}\right]\frac{x_1}{1 - x_1(a_1' + 2b\rho)} - \frac{1}{x_0}\,\frac{x_1 - x_0}{1 - (x_1 - x_0)a_1' - 2b\rho x_1}$$

$$-\frac{3}{4}a^2 b(x_1 - x_0) = 0, \tag{25}$$

with $a_1' = a_1/b$. A simple algebraic analysis of these equations shows that for a dense globular system the only real solution for x_1 is $x_1 = x_0$ and therefore $a_1 = a_2 = -2b^2\rho$. This result implies that the two-step replica symmetry breaking scheme reduces to a one-step scheme. The stability of the one-step replica symmetry breaking (RSB) solution was observed in all microscopic studies of three-dimensional heteropolymers [3–5,19]. One-step RSB suggests that the energy landscape of a polymer is ''rugged,'' consisting of well-defined local energy minima such that it is a consequence of the topology of the three-dimensional space that makes ''half-folded'' states unfavorable: loss of energy due to severing some favorable contacts is not fully compensated by entropy gain in such ''half-folded'' states. This is in contrast to low-dimensional heteropolymers where the mean-field solution features continuous RSB. Plotkin *et al.* [20] *postulated* the possibility of a continuous replica symmetry breaking in their phenomenological description of random heteropolymers based on the generalized random energy model. In contrast, our analysis based on the *microscopic* model shows that a stable mean-field solution for three-dimensional heteropolymers features one-step replica symmetry breaking.

IV. PHASE DIAGRAM BEYOND THE GAUSSIAN APPROXIMATION

In [1] terms of higher order than quadratic in the field Hamiltonian given in Eq. (11) were neglected. Within this Gaussian approximation a phase diagram was calculated. In the following we show that this phase diagram is changed qualitatively if terms of higher order in the field ϕ are included. Therefore we make the following perturbation expansion of the averaged replicated partition function [Eq. (11)]:

$$\langle Z^n\rangle_{\rm av}\approx\frac{1}{\tilde{Z}}\Bigg\langle\int\mathcal{D}\Phi_\alpha\mathcal{D}\Phi_0\exp\Bigg[-\sum_{\alpha=0}^{n}b_\alpha\int d\mathbf{R}_1 d\mathbf{R}_2\Phi_\alpha(\mathbf{R}_1)\Phi_\alpha(\mathbf{R}_2)$$
$$\times U(\mathbf{R}_1-\mathbf{R}_2)+2\sum_{\alpha,\beta=0}^{n}b_\alpha b_\beta\int d\mathbf{R}_1 d\mathbf{R}_1'\Phi_\alpha(\mathbf{R}_1')U(\mathbf{R}_1-\mathbf{R}_1')$$
$$\times\int d\mathbf{R}_2 d\mathbf{R}_2'\Phi_\alpha(\mathbf{R}_2')U(\mathbf{R}_2-\mathbf{R}_2')Q_{\alpha\beta}(\mathbf{R}_1-\mathbf{R}_2)\Bigg]$$
$$\times\Bigg[1-\frac{4}{3}\sum_{\alpha,\beta,\gamma,\delta}b_\alpha b_\beta b_\gamma b_\delta\int d\mathbf{R}_1 d\mathbf{R}_1'\cdots d\mathbf{R}_4 d\mathbf{R}_4'$$
$$\times\Phi_\alpha(\mathbf{R}_1')U(\mathbf{R}_1-\mathbf{R}_1')\cdots\Phi_\delta(\mathbf{R}_4')U(\mathbf{R}_4-\mathbf{R}_4')$$
$$\times\sum_i\delta(\mathbf{r}_i^\alpha-\mathbf{R}_1)\,\delta(\mathbf{r}_i^\beta-\mathbf{R}_2)\,\delta(\mathbf{r}_i^\gamma-\mathbf{R}_3)\,\delta(\mathbf{r}_i^\delta-\mathbf{R}_4)\Bigg]\Bigg\rangle_{\rm th}. \qquad (26)$$

Here terms of higher order than ϕ^4 were neglected. It is useful to consider this expression in Fourier space, where it has the following form:

$$\langle Z^n\rangle_{\rm av}\approx\frac{1}{\tilde{Z}}\Bigg\langle\int\mathcal{D}\Phi_\alpha\mathcal{D}\Phi_0\exp\Bigg[-\sum_{\alpha=0}^{n}\sum_{\mathbf{k}}b_\alpha\Phi_\alpha(\mathbf{k})\Phi_\alpha(-\mathbf{k})+2\sum_{\alpha,\beta=0}^{n}b_\alpha b_\beta$$
$$\times\sum_{\mathbf{k}}Q_{\alpha\beta}(\mathbf{k})\Phi_\alpha(\mathbf{k})\Phi_\beta(-\mathbf{k})\Bigg]\Bigg[1-\frac{4}{3}\sum_{\alpha,\beta,\gamma,\delta}b_\alpha b_\beta b_\gamma b_\delta\int d\mathbf{k}_1 d\mathbf{k}_2 d\mathbf{k}_3 \mathrm{d}\mathbf{k}_4$$
$$\times\sum_i\Phi_\alpha(\mathbf{k}_1)e^{-i\mathbf{k}_1\mathbf{r}_i^\alpha}\Phi_\beta(\mathbf{k}_2)e^{-i\mathbf{k}_2\mathbf{r}_i^\beta}\Phi_\gamma(\mathbf{k}_3)e^{-i\mathbf{k}_3\mathbf{r}_i^\gamma}\Phi_\delta(\mathbf{k}_4)e^{-i\mathbf{k}_4\mathbf{r}_i^\delta}\Bigg]\Bigg\rangle_{\rm th}. \qquad (27)$$

The first integral in Eq. (27) is a gaussian integral in the fields $\{\phi\}$. It has already been treated in [10] i.e., in [1]. A one-step replica symmetry breaking scheme was applied to this problem, which is valid as we pointed out in the previous paragraph. The replica symmetry breaking parameter is referred to as x_0 and within this scheme for the unperturbated free energy one gets [10]

$$C(x_0)=\ln(b)+\frac{\ln(1-2b\rho x_0)}{x_0}-\frac{s}{x_0}, \qquad (28)$$

where ρ is the density of the system and s is the flexibility parameter or entropy per monomer defined by $s=\ln(a^3/v)$ with a^3 the volume of a monomer and v the excluded volume. So a^3/v is the number of possibilities to place a certain monomer along a given chain structure to achieve coincidence with this structure on a microscopic level. For further details on the motivation of the definition of s see [10].

In the second integral in Eq. (27) only terms that contain pairwise replica indices from the same group will survive, since $Q_{\alpha\beta}(\mathbf{k})$ vanishes for α and β belonging to different groups. This leads to

$$\langle Z^n\rangle_{\rm av}\approx C(x_0)-C(x_0)\Bigg\langle\frac{16}{3}\sum_{\mathbf{k}_1,\mathbf{k}_2}\sum_{(A,B)}\sum_{\alpha,\beta\epsilon A;\gamma,\delta\epsilon B}\sum_i b_\alpha b_\beta b_\gamma b_\delta e^{i\mathbf{k}_1(\mathbf{r}_i^\alpha-\mathbf{r}_i^\beta)}e^{i\mathbf{k}_2(\mathbf{r}_i^\gamma-\mathbf{r}_i^\delta)}[P^{-1}]^A_{\alpha\beta}(\mathbf{k}_1)[P^{-1}]^B_{\gamma\delta}(\mathbf{k}_2)\Bigg\rangle_{\rm th}$$
$$-C(x_0)\Bigg\langle\frac{16}{3}\sum_{\mathbf{k}_1,\mathbf{k}_2}\sum_{(A)}\sum_{\alpha,\beta\epsilon A;\alpha,\beta\neq0}\sum_i b_\alpha b_\beta b_s^2 e^{i\mathbf{k}_1(\mathbf{r}_i^\alpha-\mathbf{r}_i^\beta)}e^{i\mathbf{k}_2(\mathbf{r}_i^0-\mathbf{r}_i^0)}[P^{-1}]^A_{\alpha\beta}(\mathbf{k}_1)\frac{1}{b_s-2b_s^2\rho}\Bigg\rangle_{\rm th}, \qquad (29)$$

with A and B denoting the different groups of replicas. $P^{-1}(\mathbf{k})$ is the inverse Parisi matrix represented by $p(\mathbf{k})$ for the off-diagonal elements and $\tilde{p}(\mathbf{k})$ for the diagonal elements:

$$p(\mathbf{k})=\frac{\gamma(\mathbf{k})}{b[1-\gamma(\mathbf{k})x_0]} \qquad (30)$$

and

$$\tilde{p}(\mathbf{k})=\frac{1+\gamma(\mathbf{k})(1-x_0)}{b[1-\gamma(\mathbf{k})x_0]} \tag{31}$$

with $\gamma(\mathbf{k})=2b\rho$. Suppose that b and γ do not depend on $\mathbf{k}$. Then

$$\begin{aligned}\langle Z^n\rangle_{\rm av}\approx{}& C(x_0)-\frac{16}{3}C(x_0)\sum_{\mathbf{k}_1,\mathbf{k}_2} b^4\frac{n}{x_0}\left(\frac{n}{x_0}-1\right)[(x_0-1)x_0p+x_0\tilde{p}]^2\\ &\times\left\langle\sum_i e^{i\mathbf{k}_1(\mathbf{r}_i^\alpha-\mathbf{r}_i^\alpha)}e^{i\mathbf{k}_2(\mathbf{r}_i^\gamma-\mathbf{r}_i^\gamma)}\right\rangle_{\rm th}-\frac{16}{3}C(x_0)\sum_{\mathbf{k}_1,\mathbf{k}_2} b^4\frac{n}{x_0}\\ &\times[3x_0(x_0-1)(x_0-2)(x_0-3)p^2+x_0(x_0-1)(x_0-2)(2p^2+p\tilde{p})\\ &+3x_0(x_0-1)p\tilde{p}+3x_0\tilde{p}^2+x_0(x_0-1)(2p^2+\tilde{p}^2)]\\ &\times\left\langle\sum_i e^{i\mathbf{k}_1(\mathbf{r}_i^\alpha-\mathbf{r}_i^\alpha)}e^{i\mathbf{k}_2(\mathbf{r}_i^\alpha-\mathbf{r}_i^\alpha)}\right\rangle_{\rm th}-\frac{16}{3}C(x_0)\frac{n}{x_0}\sum_{\mathbf{k}_1,\mathbf{k}_2}[(x_0-1)x_0p+x_0\tilde{p}]\frac{b_s^2b^2}{b_s-2b_s^2\rho}\\ &\times\left\langle\sum_i e^{i\mathbf{k}_1(\mathbf{r}_i^\alpha-\mathbf{r}_i^\alpha)}e^{i\mathbf{k}_2(\mathbf{r}_i^0-\mathbf{r}_i^0)}\right\rangle_{\rm th}.\end{aligned} \tag{32}$$

Therefore

$$\begin{aligned}\langle Z^n\rangle_{\rm av}\approx{}& C(x_0)-\frac{4\gamma^2C(x_0)(n-x_0)n}{3(1-\gamma x_0)^2}-\frac{4C(x_0)\gamma\gamma_s n}{3(1-\gamma x_0)(1-\gamma_s)}\\ &-\frac{4\gamma^2C(x_0)n[3\gamma^2(x_0^3-5x_0^2+9x_0-3)-\gamma(x_0^2+4x_0-3)+x_0+2]}{3(1-\gamma x_0)^2},\end{aligned} \tag{33}$$

which leads to the following free energy density of the system:

$$\begin{aligned}\frac{f}{n}={}& C(x_0)-\frac{4\gamma^2C(x_0)(n-x_0)}{3(1-\gamma x_0)^2}-\frac{4C(x_0)\gamma\gamma_s}{3(1-\gamma x_0)(1-\gamma_s)}\\ &-\frac{4\gamma^2C(x_0)[3\gamma^2(x_0^3-5x_0^2+9x_0-3)-\gamma(x_0^2+4x_0-3)+x_0+2]}{3(1-\gamma x_0)^2}.\end{aligned} \tag{34}$$

Fluctuations of the order parameter $Q_{0\alpha}$, which describes the overlap of the αth replica with the target state, might affect the free energy density. This effect was investigated in [1], where bilinear terms $\phi_\alpha(\mathbf{k})\phi_0(-\mathbf{k})$ in the field theory were taken into account. The result of this consideration is that the correction of the free energy density due to the fluctuations reads

$$\left(\frac{f}{n}\right)_{\rm cor}=\frac{\epsilon_0\gamma\gamma_s}{(1-\gamma x_0)(1-\gamma_s)}. \tag{35}$$

ϵ_0 is a small parameter of a perturbation expansion defined by

$$\epsilon_0=\frac{1}{N}\left\langle\sum_{i,j}\delta(\mathbf{r}_i^\alpha-\mathbf{r}_j^\alpha)\,\delta(\mathbf{r}_i^0-\mathbf{r}_j^0)\right\rangle_{\rm th}. \tag{36}$$

The thermal expectation value in Eq. (36) is the number of contacts different folds α and the target configuration have in common. The overlap is mainly due to the contacts of neighboring monomers which is neglected in the mean-field theory for $Q_{0\alpha}$ and might become important when the flexibility of the chain increases. The overall free energy is the sum of Eqs. (35) and (34)

$$\frac{f}{n}=C(x_0)-\frac{4\gamma^2C(x_0)(n-x_0)}{3(1-\gamma x_0)^2}-\frac{4C(x_0)\gamma\gamma_s}{3(1-\gamma x_0)(1-\gamma_s)}$$
$$-\frac{4\gamma^2C(x_0)[3\gamma^2(x_0^3-5x_0^2+9x_0-3)-\gamma(x_0^2+4x_0-3)+x_0+2]}{3(1-\gamma x_0)^2}+\frac{\epsilon_0\gamma\gamma_s}{(1-\gamma x_0)(1-\gamma_s)}. \tag{37}$$

This expression for the free energy has to be maximized with respect to the one-step replica symmetry breaking parameter x_0. Therefore we calculate the partial derivative and set the result to zero:

$$\frac{\partial f/n}{\partial x_0}=C'(x_0)-\frac{4\gamma^2C'(x_0)x_0}{3(1-\gamma x_0)^2}-\frac{4\gamma^2C(x_0)}{3}\left[\frac{1}{(1-\gamma x_0)^2}+\frac{2\gamma x_0}{(1-\gamma x_0)^3}\right]-\frac{4C'(x_0)\gamma\gamma_s}{3(1-\gamma x_0)(1-\gamma_s)}-\frac{4C(x_0)\gamma^2\gamma_s}{3(1-\gamma x_0)^2(1-\gamma_s)}$$
$$+\frac{\epsilon_0\gamma^2\gamma_s}{(1-\gamma x_0)^2(1-\gamma_s)}-\frac{4\gamma^2C'(x_0)[3\gamma^2(x_0^3-5x_0^2+9x_0-3)-\gamma(x_0^2+4x_0-3)+x_0+2]}{3(1-\gamma x_0)^2}$$
$$+\frac{4\gamma^2C(x_0)[3\gamma^3(x_0^3-9x_0+6)-\gamma^2(9x_0^2-34x_0+33)+\gamma x_0-1]}{3(1-\gamma x_0)^3}=0, \tag{38}$$

with $C'(x_0)=\partial C(x_0)/\partial x_0$. Due to the logarithmic terms in $C(x_0)$, this is a transcendent equation. So the general solution for x_0 is unknown. For our purpose, however, we do not need to know the solution for x_0, since we are mainly interested in calculating the temperature T_c at which a transition from the disordered phase to the frozen phase of the considered system occurs. At this freezing transition the replicas start to form groups, which means that x_0 becomes smaller than 1. Consequently at the transition point we get $x_0(T_c)=1$. Considering Eq. (38) at $T=T_c$ for small interaction parameters χ (i.e., $\chi/T<<1$) yields

$$s-\frac{2\chi^2\rho^2}{T_c^2}-\frac{32\chi^2}{3T_c^2}s-\frac{32\chi^2\rho^2}{3T_c^2}\ln\left(\frac{-\chi}{T_c}\right)$$
$$-\frac{16\chi^2\rho^2 s}{3T_cT_s}+\mathcal{O}(\chi^3)=0. \tag{39}$$

The restriction to only small interactions is consistent with the fact that terms of higher order in the field theory are neglected. The analysis of Eq. (39) clearly shows that the critical temperature T_c increases for decreasing selective temperatures T_s, which is in contrast to the dependence of T_c on T_s calculated in [1]. Though Eq. (39) cannot be solved analytically, it can be solved numerically in order to plot a phase diagram.

It is interesting to consider the limit of a high selective temperature T_s in Eq. (39), which represents the case of a random copolymer. It is well known that random copolymers show a finite freezing temperature T_c^r, which was calculated by Sfatos *et al.* [10] with a similar formalism than presented in this work. Like in [1] Sfatos *et al.* [10] made a quadratic approximation in the field ϕ as pointed out above. It can be seen from Eq. (39) that considering infinite selective temperatures T_s gives a finite $T_c^r>0$ depending on the interaction parameter χ, the density ρ, and the flexibility parameter s.

It is important to mention that Eq. (39) only delivers a solution for the transition temperature T_c if $-\chi/T_c$ is not too small. On the other hand, our theory only is valid if $-\chi/T_c$ is not too big i.e., at least smaller than 1. It is, however, easy to show that Eq. (39) gives a solution for a wide range of $-\chi/T_c$. Equation (39) is equivalent to

$$\frac{-\chi}{T_c}=\exp\left(\frac{3sT_c^2}{32\chi^2\rho}-\frac{6}{32}-\frac{s}{\rho^2}-\frac{sT_c}{2T_s}\right). \tag{40}$$

To ensure that $-\chi/T_c$ is smaller than 1, which is necessary since the theory is based on an expansion with respect to $-\chi/T_c$, the argument of the exponential function in Eq. (40) has to be negative. This implies

$$\frac{3sT_c^2}{32\chi^2\rho}<\frac{6}{32}+\frac{s}{\rho^2}+\frac{sT_c}{2T_s}. \tag{41}$$

All in all we get

$$1>\frac{\chi^2}{T_c^2}>\frac{3s}{32s+6\rho^2+16\rho^2 sT_c/T_s}, \tag{42}$$

which is satisfied in a wide range of $-\chi/T_c$ even for small s, since the right hand side vanishes as s goes to zero.

In order to complete the phase diagram of our considered system we have to compare the free energy in the different phases. In [1] the free energy density of the native state was calculated as

$$F=-2b\rho/(1-2b_s\rho)-s,$$

$$T_s>-2\chi\rho=-\frac{3}{16}b/(\rho b_s^2)[1-1/(2b_s\rho)]-s,$$

$$T_s<-2\chi\rho. \tag{43}$$

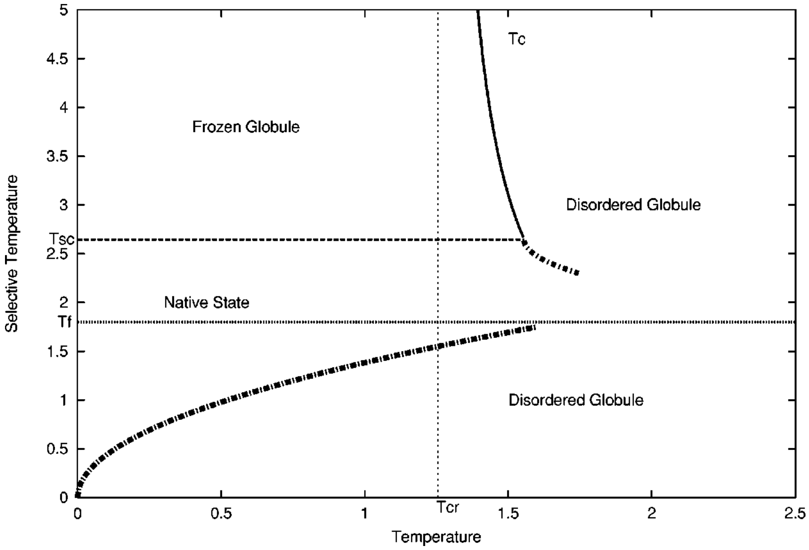

FIG. 1. Phase diagram for a heteropolymeric chain with a selected sequence with the parameters of real temperature and selective or polymerization temperature in arbitrary units.

A phase transition in the sequence space of the monomers on the inhomogeneous target conformation was observed at $T_s = T_f = -2\chi\rho$, which is analogous to the ferromagnetic-paramagnetic transition in a ferromagnet [1]. This means that the two kinds of monomers A and B form domains of equal monomers when synthezised. If the protein is for example exposed to water, it will establish a hydrophobic core and a hydrophilic surface below T_f.

The comparison of the free energy density of the native state above T_f, Eq. (43), and of the frozen globule, Eq. (37) at $T = T_c$, which implies $x_0 = 1$, gives the transition temperature in sequence space T_s^c between the native state and the frozen globule. Note that T_s^c, which can be calculated numerically, is independent of the temperature in real space, since the entropy of both the native state and frozen globule vanishes.

The transition line between the disordered globule and the native state can be calculated in the neighborhood of the frozen globule, which means that the free energy densities of the native state above T_f, Eq. (43) and of the disordered globule, Eq. (37), for $x_0 = 1$ are equal.

Below the temperature T_f the free energy density of the disordered globule can be obtained to be [1]

$$f = -\frac{3}{16\rho^2}\left(\frac{1}{b} - 2\rho\right)^2. \tag{44}$$

Comparing this free energy density with the one for the native state below the sequence space temperature T_f one obtains a transition line for low real temperatures and sequence space temperatures defined by [1]

$$T \propto T_s^2. \tag{45}$$

The results are summarized in the phase diagram given in Fig. 1.

The line denoted by T_s^c is the transition line from the frozen globule to the native state, whereas the T_c line represents the transition between the disordered globule and the frozen globule above T_s^c. The part of this line below T_s^c indicates a first-order transition to the native state. This transition is of first order since there is a entropy difference between the disordered globule and the single native state. Another first-order transition [1] occurs between the disordered globule and the native state below the line denoted by T_f. This T_f line represents the ferromagnetic like transition within the native state in sequence space. The dotted vertical line labeled T_c^r is the asymptote for the T_c transition line in the limit of high selective temperatures T_s, which creates a random copolymer.

This phase diagram is qualitatively different in one aspect from the phase diagram presented in [1], which was calculated within the Gaussian approximation of the field Hamiltonian. The major difference is that due to the ϕ^4 correction in the field Hamiltonian the freezing temperature T_c for the transition between the disordered and the frozen globule decreases with an increasing selective temperature, whereas within the Gaussian approximation [1] T_c is slightly increasing with an increasing selective temperature. This is in accordance with the intuitive understanding which suggests that the freezing temperature increases as the selective temperature decreases, since the sequence becomes more ordered by lowering the selective temperature. We note that the very weak dependence of T_c on selective temperature observed in [1] was due to fluctuations in the order parameter $Q_{\alpha\beta}$ that physically take into account the fact that all conformations share same set of local (along the chain) contacts (the relative contribution of local contacts is given by parameter ϵ^0). The slight decrease of T_c at lower selective temperature was due to the fact that sequence selection induced some short-range correlation, making the chain more homopolymer like. We note that neglect of fluctuations of $Q_{\alpha\beta}$ in the Gaussian approximation (i.e., setting $\epsilon^0 = 0$) leads to the independence of T_c on selective temperature. The reason for this is simple: in the Gaussian approximation optimization of sequences to fit the target conformation does not induce any correlations in sequences $T > T_f$ other than related to contacts present in the target conformation (see below). Since the structural overlap between the native state and any of the frozen conformations is small, in the thermodynamic limit (because of the one-step RSB solution for $Q_{0\alpha}$) the sequences, in the Gaussian approximation, are effectively random from the point of view of freezing into random conformations: hence, the independence of T_c on selective temperature, apart from local correlation effects (governed by ϵ^0), in this approximation. When higher-order terms in ϕ are included, the situation changes as sequence fluctuations are treated more consistently. (This is usual—departure from the Gaussian approximation allows us to properly take into account fluctuations.) Apparently these sequence fluctuations affect the freezing transition as seen in the phase diagram.

In the previous section it was provend that there is no two-step replica symmetry breaking in the Gaussian approximation. There, however, might be a break of the replica symmetry of higher order if one goes beyond the Gaussian approximation, which might have an effect on the phase diagram presented here. This will be investigated in future work.

V. SEQUENCE CORRELATION FUNCTION

Since the monomer sequence of the chain is selected and not random within our model, the sequence correlation function is an important quantity to study. It is defined as

$$\langle \sigma_l \sigma_{l+k} \rangle_{\rm av} = \frac{1}{\tilde{Z}} \sum_{\{\sigma_i\}} \sigma_l \sigma_{l+k} \exp\left(-\frac{\mathcal{H}_0(\{\sigma_i\})}{T_s} \right). \tag{46}$$

Here $\langle \cdots \rangle_{\rm av}$ represents the average over all sequences with a fixed conformation of the chain. This expectation value can be rewritten as

$$\langle \sigma_l \sigma_{l+k} \rangle_{\rm av} = \frac{1}{\tilde{Z}} \sum_{\{\sigma_i\}} \sigma_l \sigma_{l+k} \exp\left[b_s \int d\mathbf{R}_1 \int d\mathbf{R}_2 \sum_i \sigma_i \delta(\mathbf{r}_i^0 - \mathbf{R}_1) U(\mathbf{R}_1 - \mathbf{R}_2) \sum_j \sigma_j \delta(\mathbf{r}_j^0 - \mathbf{R}_2) \right]. \tag{47}$$

Introducing a field theory by performing a Hubbard-Stratonovich transformation the second exponential function in Eq. (47) becomes

$$\int \mathcal{D}\psi(\mathbf{R}) \exp\left[-\frac{1}{4b_s} \int d\mathbf{R}_1 \int d\mathbf{R}_2 \psi(\mathbf{R}_1) U^{-1}(\mathbf{R}_1 - \mathbf{R}_2) \psi(\mathbf{R}_2) + \int d\mathbf{R}\, \psi(\mathbf{R}) \sum_i \sigma_i \delta(\mathbf{r}_i^0 - \mathbf{R}) \right]. \tag{48}$$

Performing the trace over $\{\sigma_i\}$ and implementing Eq. (48) into Eq. (47) yields

$$\begin{aligned} \langle \sigma_l \sigma_{l+k} \rangle_{\rm av} = \frac{1}{\tilde{Z}} \int \mathcal{D}\psi(\mathbf{R}) \exp\Bigg[& -\frac{1}{4b_s} \int d\mathbf{R}_1 \int d\mathbf{R}_2 \\ & \times \psi(\mathbf{R}_1) U^{-1}(\mathbf{R}_1 - \mathbf{R}_2) \psi(\mathbf{R}_2) + \sum_{i=l,l+k} \ln\left[\sinh\left(\int d\mathbf{R}\, \psi(\mathbf{R}) \delta(\mathbf{r}_i^0 - \mathbf{R}) \right) \right] \\ & + \sum_{i \neq l,l+k} \ln\left[\cosh\left(\int d\mathbf{R} \psi(\mathbf{R}) \delta(\mathbf{r}_i^0 - \mathbf{R}) \right) \right] \Bigg]. \end{aligned} \tag{49}$$

Developing ln(sinh) and ln(cosh) up to second order in the field ψ one obtains

$$\begin{aligned} \langle \sigma_l \sigma_{l+k} \rangle_{\rm av} \approx \frac{1}{\tilde{Z}} \int \mathcal{D}\psi \int d\mathbf{R}_1 \int d\mathbf{R}_2 \psi(\mathbf{R}_1) \psi(\mathbf{R}_2) \delta(\mathbf{r}_l^0 - \mathbf{R}_1) \delta(\mathbf{r}_{l+k}^0 - \mathbf{R}_2) \exp\Bigg[& -\frac{1}{4b_s} \int d\mathbf{R}_1 \int d\mathbf{R}_2 \\ \times \psi(\mathbf{R}_1) U^{-1}(\mathbf{R}_1 - \mathbf{R}_2) \psi(\mathbf{R}_2) + \frac{1}{6} \sum_{i=l,l+k} \psi(\mathbf{R}_1) \psi(\mathbf{R}_2) \delta(\mathbf{r}_i^0 - \mathbf{R}_1) \delta(\mathbf{r}_i^0 - \mathbf{R}_2) & \\ + \frac{1}{2} \sum_{i \neq l,l+k} \psi(\mathbf{R}_1) \psi(\mathbf{R}_2) \delta(\mathbf{r}_i^0 - \mathbf{R}_1) \delta(\mathbf{r}_i^0 - \mathbf{R}_2) \Bigg]. & \end{aligned} \tag{50}$$

Including terms of higher order in the field ψ like we did in the previous section is not recommended at this point, since here it just makes things more complicated and does not give any deeper insights.

The next step is to make a field transformation from the Hubbard-Stratonovich field ψ to the field ϕ defined in Eq. (10). Furthermore, we assume that the short-ranged potential U is a δ function. Then Eq. (50) becomes

$$\langle\sigma_l\sigma_{l+k}\rangle_{\rm av}\approx\frac{1}{\tilde{\tilde{Z}}}\int\mathcal{D}\phi 4b_s^2\int d\mathbf{R}_1\int d\mathbf{R}_2\,\phi(\mathbf{R}_1)\,\phi(\mathbf{R}_2)\,\delta(\mathbf{r}_l^0-\mathbf{R}_1)\,\delta(\mathbf{r}_{l+k}^0-\mathbf{R}_2)\exp\Bigg[-b_s\int d\mathbf{R}\,\phi^2(\mathbf{R})$$
$$+\frac{2}{3}b_s^2\sum_{i=l,l+k}\int d\mathbf{R}_1\int d\mathbf{R}_2\,\phi(\mathbf{R}_1)\,\phi(\mathbf{R}_2)\,\delta(\mathbf{r}_i^0-\mathbf{R}_1)\,\delta(\mathbf{r}_i^0-\mathbf{R}_2)$$
$$+2b_s^2\sum_{i\neq l,l+k}\int d\mathbf{R}_1\int d\mathbf{R}_2\,\phi(\mathbf{R}_1)\,\phi(\mathbf{R}_2)\,\delta(\mathbf{r}_i^0-\mathbf{R}_1)\,\delta(\mathbf{r}_i^0-\mathbf{R}_2)\Bigg].\tag{51}$$

The functional integral on the right hand side of Eq. (51) only has a contribution if $\mathbf{r}_l^0=\mathbf{r}_{l+k}^0$ which means that the monomers l and $l+k$ form a native contact. In particular we get

$$\langle\sigma_l\sigma_{l+k}\rangle_{\rm av}=\frac{2b_s\sqrt{1-4b_s}}{\left(1-\frac{4}{3}b_s\right)^{(3/2)}}\,\delta(\mathbf{r}_l^0-\mathbf{r}_{l+k}^0).\tag{52}$$

For a weakly interacting system, which means small b_s, this equation reduces to

$$\langle\sigma_l\sigma_{l+k}\rangle_{\rm av}=2b_s\,\delta(\mathbf{r}_l^0-\mathbf{r}_{l+k}^0)=-\frac{2\chi}{T_s}\,\delta(\mathbf{r}_l^0-\mathbf{r}_{l+k}^0).\tag{53}$$

This indicates that there is a correlation between the monomers that form a native contact, which is proportional to the strength of interaction and inversely proportional to the temperature in sequence space. Note that Eq. (52) is only valid for a sufficiently small interaction parameter b_s, since our theory breaks down for too strong monomer-monomer interactions.

VI. CONCLUSION

In the present paper we studied the phase diagram of a two-letter heteropolymer with selected sequences, which is a good analytical model for proteins. The same major phases as in [1], [21], and [22]—disordered, frozen, and target states—were found.

The sequence of the monomers was treated as a frozen disorder. To deal with this problem the replica trick was employed. We showed that a one-step replica symmetry breaking scheme like was applied in [1] is appropriate to solve the problem. The physical meaning of this result is that there exists a sharp transition into the native state.

Furthermore, to calculate the phase diagram we went beyond a Gaussian approximation, which resulted in a qualitative change of the phase diagram compared to that given in [1]. As presented in the phase diagram the transition between the disordered phase and the native state is thermodynamically of first order, which is due to the selection of the monomers. For higher selective temperatures $T_s>T_s^c$ the selection is much weaker and we get a transition of second order. This result is consistent with experiments which show that the formation of the molten globular state occurs as a first-order transition [23,24] in contrast to the behavior of random sequences [25]. As was mentioned in [1] for the kinetical accessibility of the native state it is important that this state be accessible from the disordered state for temperatures above the freezing temperature T_c^r for a random copolymer. In the phase diagram presented in [1] the native state is accessible in this sense, though it is required that the selective temperature be a certain amount smaller than T_s^c. The calculations in this paper, however, show that the native state is even accessible for all selective temperatures below T_s^c.

The phase diagram also shows a transition line T_f, which represents a transition in sequence space [1]. This transition is analogous to a ferromagnetic-paramagnetic transition. Below a certain temperature T_f in sequence space the two kinds of monomers tend to form domains. For selective temperatures $T_s<T_f$ there is another transition line between the disordered and the native state. This transition line suggests that the smaller the selective temperature, the smaller the transition temperature, which can be explained by the fact that the sequence forms domains below T_f, so that the sequence design process is disturbed.

Moreover, a sequence correlation function along the chain depending on the selective temperature in sequence space and the strength of the interaction between the monomers was calculated. The result suggests that in the sequence design scheme employed here special emphasis is laid on the native contacts. It also has practical applications for finding the native contacts in a protein. Our sequence correlation function implies that those pairs of monomers are candidates for forming a native contact which have a peak in this correlation function calculated within a superfamily of proteins. Therefore it can be confirmed that mutations at monomer sites which form a native contact are correlated [26]. The sequence correlation function can be used for practical applications in protein analysis. In a future work we are going to study our result on real proteins [27].

ACKNOWLEDGMENTS

This work has been made possible by the support of a grant given by the German Academic Exchange Service in connection with the University Program III by the Federal and State Governments of the Federal Republic of Germany.

[1] S. Ramanathan and E. Shakhnovich, Phys. Rev. E **50**, 1303 (1994).
[2] J.D. Bryngelson and P.G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).
[3] E.I. Shakhnovich and A.M. Gutin, J. Phys. A **22**, 1647 (1989).
[4] E.I. Shakhnovich and A.M. Gutin, Biophys. Chem. **34**, 187 (1989).
[5] E.I. Shakhnovich and A.M. Gutin, Europhys. Lett. **8**, 327 (1989).
[6] T. Garel and H. Orland, Europhys. Lett. **6**, 597 (1988).
[7] E.I. Shakhnovich and A.M. Gutin, J. Phys. (Paris) **50**, 1843 (1989).
[8] G. Frederickson and S. Milner, Phys. Rev. Lett. **67**, 835 (1991).
[9] S. Stepanow, M. Schulz, and J.-U. Sommer, Europhys. Lett. **19**, 273 (1992).
[10] C. Sfatos, A.M. Gutin, and E.I. Shakhnovich, Phys. Rev. E **48**, 465 (1993).
[11] A.M. Gutin and E.I. Shakhnovich, J. Chem. Phys. **98**, 8174 (1993).
[12] Y.H. Taketomi and N. Go, Int. J. Pept. Protein Res. **7**, 445 (1975).
[13] N. Go and H. Abe, Biopolymers **20**, 991 (1981).
[14] E. Shakhnovich and A. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195 (1993).
[15] S. Obukhov, J. Phys. A **19**, 3655 (1986).
[16] I. Lifshitz, A. Yu Grosberg, and A.R. Khokhlov, Rev. Mod. Phys. **50**, 683 (1978).
[17] K. Binder and A. Young, Rev. Mod. Phys. **58**, 801 (1986).
[18] M. Mezard and G. Parisi, J. Phys. I **1**, 809 (1991).
[19] V. Pande, A.Yu. Grosberg, and T. Tanaka, Rev. Mod. Phys. **72**, 259 (2000).
[20] S. Plotkin, J. Wang, and P. Wolynes, Phys. Rev. E **53**, 6271 (1996).
[21] M. Sasai and P. Wolynes, Phys. Rev. Lett. **65**, 2740 (1990).
[22] M. Sasai and P. Wolynes, Phys. Rev. A **46**, 7979 (1992).
[23] V. Uversky, G. Semisotnov, R. Pain, and O. Ptitsyn, FEBS Lett. **314**, 89 (1994).
[24] A. Gittis, W. Stites, and E.E. Lattman, J. Mol. Biol. **232**, 718 (1993).
[25] A. Chaffotte *et al.*, Biochemistry **30**, 8067 (1991).
[26] A.R. Ortis, A. Kolinski, and J. Skolnick, J. Mol. Biol. **277**, 419 (1998).
[27] J. Wilder and E. I. Shakhnovich (unpublished).