JOURNAL OF CHEMICAL PHYSICS VOLUME 118, NUMBER 19 15 MAY 2003

# Phase behavior of a lattice protein model

Nicolas Combe[a] and Daan Frenkel[b]
*FOM Institute for Atomic and Molecular Physics, Kruislaan 407, 1098 SJ Amsterdam, The Netherlands*

(Received 4 December 2002; accepted 20 February 2003)

We present a numerical simulation of the phase behavior of a simple model for a protein solution. We find that this system can occur in three phases, namely a dilute liquid, a dense liquid and a crystal. The transition from dilute-liquid to dense-liquid takes place in the regime where the fluid phase is metastable with respect to the crystal. We have computed the relative stabilities of different crystal morphologies. In addition, we have analyzed the ''nucleation'' of the native state of an isolated lattice protein. Using a ''Gō'' model [N. Gō, J. Stat. Phys. **30**, 413 (1983)] to describe the protein, we show that a first order transition exists between the native and the coil state. We show this by analyzing the free energy barrier for the coil-to-native transition. © *2003 American Institute of Physics.* [DOI: 10.1063/1.1567256]

## I. INTRODUCTION

Many diseases, such as prion diseases (Creutzfeldt-Jakob),[1,2] Alzheimer disease[3,4] or cataract[5] are thought to be partly due to the abnormal aggregation of proteins. Aggregation is also a serious problem in many other domains such as the pharmaceutical[6] and food industries.[7] Understanding and controlling this process is thus of prime importance. A first attempt to model the phase behavior of protein-like chain molecules has been reported by Gupta *et al.*[8] However, it is fair to say that our understanding of aggregation is still far from complete.

A protein is composed of a chain of several dozens to a few thousands of amino-acids[9] and there are 20 different types of amino acids. This makes the number of possible sequences huge. Only a small fraction of all possible sequences occur in nature. The biological function of a protein depends on its ground state conformation; a protein or, more generally, a heteropolymer can have many conformations. In poor solvent conditions, a protein folds into a unique conformation which depends only on the sequence of amino acids: the native state. In contrast, most heteropolymers do not have a unique native state. The native state of a protein is the conformation that has the lowest free energy. In lattice models of proteins, the native state is the conformation with the lowest potential energy.

As fully atomistic simulations of proteins are very time-consuming, many numerical studies of proteins make use of coarse-grained models. Among these, the Gō model[10] is often used because it is very simple, yet retains the main aspects of the protein folding.[11] An alternative description is based on the so-called HP model, in which amino acids are considered to be of two types only: hydrophobic (H) and polar (P). For the description of the folding of small proteins, the Gō model has been shown to reproduce the qualitative behavior significantly better than the HP model.[12]

In the present paper, we present a thermodynamic study of both the folding of proteins and of the phase behavior of proteins in the scope of the cubic lattice Gō model. After describing the model and the different numerical techniques that we used, we present the folding behavior of isolated chains as a function of the denaturant concentration. In the second part, we present a study of the phase behavior of a multiple-protein system.

## II. MODEL

A protein is modeled by a self-avoiding chain of length $l_{\mathrm{seq}}$ on a cubic lattice. The Hamiltonian of a system containing one or several proteins is

$$H=\sum_k \sum_{i_k>j_k+1} \epsilon_{i_k j_k}\sigma_{i_k j_k}+\frac{1}{2}\sum_{k\neq k'} \sum_{i_k,j_{k'}} \epsilon_{i_k j_{k'}}\sigma_{i_k j_{k'}}. \qquad (1)$$

Amino acids are labeled according to their position in the sequence: amino acid $i_k$ is the $i$th amino acids of the protein $k$. The first term in Eq. (1) refers to intramolecular interactions; only interactions between nonconsecutive amino acids in the chain are taken into account. The second term in Eq. (1) deals with intermolecular interactions. $\sigma_{i_k j_{k'}}=1$ if amino acids $i_k$ and $j_{k'}$ are neighbors on the lattice, 0 otherwise. $[\epsilon_{ij}]$ is the interaction matrix which gives the interaction energy between amino acid number $i$ and $j$. The Gō model specifies the interaction energies between different amino acids of the same protein in such a way that the native state is uniquely favored: native contacts have an attractive interaction energy $\epsilon<0$, whereas all others possible contacts have no interaction. Concerning the intermolecular interactions, we assume that they are identical to those of individual proteins; residues that attract inside a protein also attract if they do not belong to the same proteins and moreover, we assume that equivalent residues in different proteins also attract with an interaction energy $\epsilon$: $\epsilon_{ii}=\epsilon$ for $1\leqslant i\leqslant l_{\mathrm{seq}}$.

Calling $n$ the number of intra- or intermolecular bonds, the partition function of the system has a very simple form,

$$Z=\sum_{\mathrm{config}} e^{-n\epsilon/k_bT}. \qquad (2)$$

[a] Electronic mail: combe@amolf.nl
[b] Electronic mail: frenkel@amolf.nl

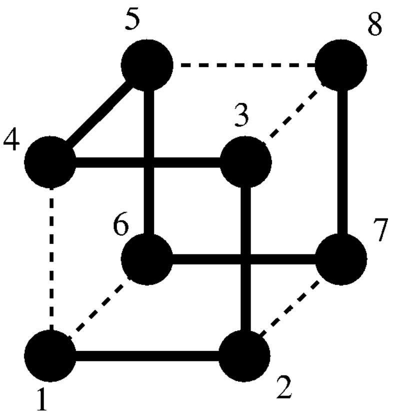

FIG. 1. Three-dimensional representation of the native structure of a short model peptide. Native contacts are denoted by dashed lines.

Here and in the following, "bond" will only refer to amino acids that are neighbors on the lattice with a nonzero interaction energy.

Figure 1 shows the native state of one of the proteins that we study. This short chain of length 8 will be used in Sec. IV to evaluate the phase diagram. To be more explicit about the model, the interaction matrix deduced from that native state is

$$[\epsilon_{ij}]=\begin{bmatrix} \epsilon & 0 & 0 & \epsilon & 0 & \epsilon & 0 & 0 \\ 0 & \epsilon & 0 & 0 & 0 & 0 & \epsilon & 0 \\ 0 & 0 & \epsilon & 0 & 0 & 0 & 0 & \epsilon \\ \epsilon & 0 & 0 & \epsilon & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \epsilon & 0 & 0 & \epsilon \\ \epsilon & 0 & 0 & 0 & 0 & \epsilon & 0 & 0 \\ 0 & \epsilon & 0 & 0 & 0 & 0 & \epsilon & 0 \\ 0 & 0 & \epsilon & 0 & \epsilon & 0 & 0 & \epsilon \end{bmatrix}$$

In our simulations, we have used a number of different computational techniques. In Sec. III, for the chain of length 8, we have enumerated explicitly all the conformations of the chain. This allows exact calculation of all thermodynamical quantities of isolated chains. For longer chains and for the work of Sec. IV, we can no longer generate all possible configurations of the system. In that case, we use Monte Carlo simulations in order to sample the most relevant part of the phase space.

In the Monte Carlo simulations, we used both "local" moves (corner move, crankshaft move, end move, reptation)[13–15] and global moves using the Configurational-Bias Monte Carlo algorithm[16] to generate different conformations of the proteins and thus, different configurations of the system. Global Monte Carlo moves yield a good acceptance rate at low densities (for the case of a multiprotein system and especially for short proteins), but local moves are more efficient at high densities, and for low temperature simulations of long isolated proteins. We stress that our aim is to explore the thermodynamics (in particular, the phase behavior) of protein model systems. Hence, the (lack of) realism of the dynamics generated by our Monte Carlo moves is less relevant.

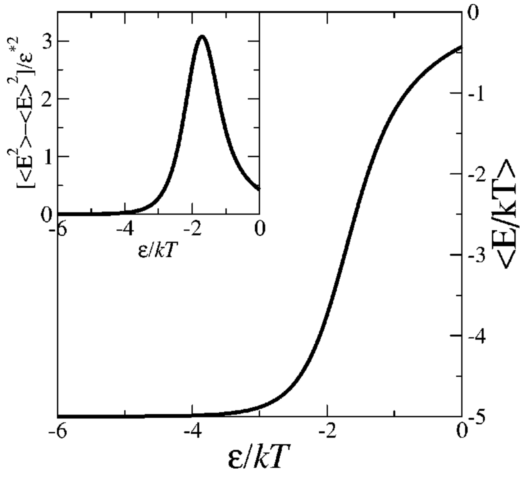

FIG. 2. Average conformational energy of the protein given in Fig. 1 as a function of the reduced energy $\epsilon/k_bT$. In the inset, the standard deviation of the energy is presented as a function of the reduced quantity.

## III. PHASE BEHAVIOR OF ISOLATED PROTEINS

Before proceeding to simulations of a multichain system, we first analyze the behavior of an isolated protein. As in homopolymers, a transition from a coil state to the native state can be induced by changing $\epsilon$. Figure 2 shows the average and the standard deviation (proportional to the heat capacity) of the energy of an isolated protein, as a function of $\epsilon/k_bT$. These curves have been obtained by explicit computation of all possible conformations of the protein.

Figure 2 shows the transition between the native state and the coil state. The maximum of the heat capacity provides an indication of the transition temperature, $\epsilon_t/k_bT=-1.71$. We stress, however, that the transition is not sharp. The heat capacity curve shows that only two states of the chains are present: the native and the coil state. The coil state consists of free chains with almost no intramolecular interactions. More specifically, Fig. 2 does not present any molten globule state. This is probably due to the fact that we study extremely short lattice proteins. Longer lattice proteins are expected to exhibit a molten globule state, as in real proteins.[17]

To further investigate the nature of the transition between the two observed states, we calculate the free energy of the system as a function of the number of native bonds. More precisely, we define a reduced partition function and a reduced free energy depending on the number of native bond $n_0$,

$$Z(n_0)=\sum_{\text{config}}\delta(n-n_0)\exp(-n\epsilon/k_bT), \tag{3}$$

$$F(n_0)=-k_bT\ln Z(n_0), \tag{4}$$

where $n$ is the number of native bonds and $\delta(n-n_0)=1$ if $n=n_0$, and 0 otherwise.

Figure 3 shows the free energy divided by $\epsilon$ as a function of the number of bonds for different values of $\epsilon/k_bT$ and for the chain of Fig. 1. As expected and in agreement with Fig. 2, the coil state is the most stable state for high values of

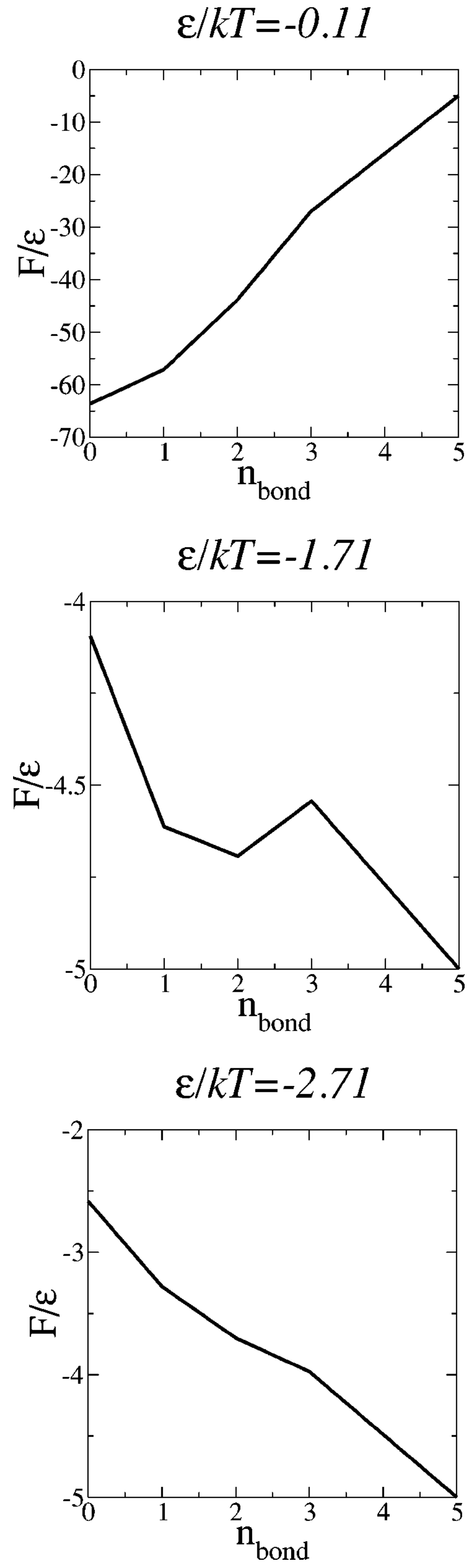

FIG. 3. Conformational free energy (in unit of $\epsilon$) of the chain of length 8 as a function of the number of native bonds. The plot is given for three different values of $\epsilon/k_bT$.

$\epsilon/k_bT$, whereas the native state is more stable for low values of $\epsilon/k_bT$. At the transition, the free energy landscape exhibits a free energy barrier.

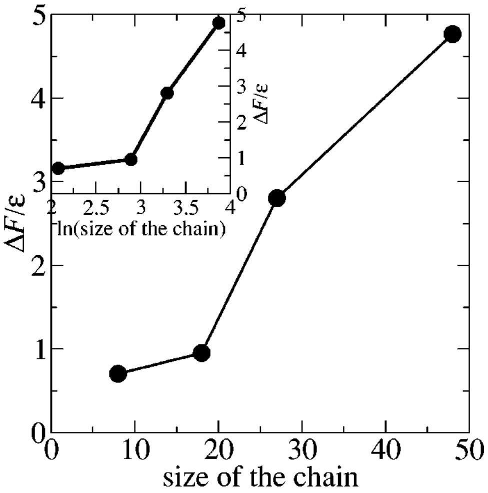

FIG. 4. Height of the free energy barrier (in units of $\epsilon$) at the coexistence between the coil and the native state as a function of the protein size. In the inset, the height of the free energy barrier (in units of $\epsilon$) at coexistence is plotted as a function of the logarithm of the protein size.

The Gō model thus shows a two state behavior which is in agreement with experimental observations of short proteins.[18]

To determine the order of that transition, we have calculated the height of the free energy barrier as a function of the size of the chains. We evaluate the free energy barriers for chains of 18, 27, and 48 amino acids. The native states of these proteins are similar to the one of the short chain already mentioned, i.e., they are fully compact with rectangular parallelepipedal (18 and 48) or cubic (27) shapes. The coexistence of the two states can be determined either from the heat capacity curve or by equalizing the probability of being in each basin of attraction defined by Fig. 3(b). More precisely, we define the transition state as the local maximum of the free energy curve [see Fig. 3(b)], and we call $n_{\text{trans}}$ the number of bonds in this state [$n_{\text{trans}}=3$ in the case of Fig. 3(b)]. The coexistence between the coil and native state is then given by the equality of the probabilities to be in each basin,

$$\frac{\sum_{i=0}^{n_{\text{trans}}} Z(i)}{\sum_{i=0}^{n_{\text{native}}} Z(i)} = \frac{\sum_{i=n_{\text{trans}}}^{n_{\text{native}}} Z(i)}{\sum_{i=0}^{n_{\text{native}}} Z(i)}, \tag{5}$$

where $n_{\text{native}}$ is the number of bonds in the native state and $Z(i)$ is given by Eq. (3). We then define the free energy barrier from the ratio of the probability to be in the transition state on the probability to be in one of the basin,

$$e^{-\beta\Delta F} = \frac{Z(n_{\text{trans}})}{\sum_{i=0}^{n_{\text{trans}}} Z(i)}. \tag{6}$$

Figure 4 shows the variation of $\Delta F/\epsilon$ as a function of the size of the proteins. This plot shows clearly that the free energy barrier increases as the size of the chains grows. A linear regression of the curve shows that the free energy barrier increases as $0.11\pm 0.01$ times the size of the proteins. If we assume that observed chain-size dependence persists for longer chains, this observation suggests that, in the Gō model, the coil-native transition is first order.

As can be seen in Fig. 3, the free energy barrier for folding depends on temperature. Dynamical simulations by Gutin *et al.*[19] show that the minimum folding time as a function of the temperature increases as a power function of the size of the chains. However, these results cannot be compared directly with our data for the temperature dependence of the free-energy barrier, as folding rates depend both on the barrier height and on a kinetic prefactor. Nevertheless, if we assume that the folding time has a simple, Arrhenius-type dependence on the folding barrier, then we can determine an ''effective'' power law that describes the dependence of the minimum folding time on chain size. In doing so, we obtain an effective exponent λ between 2.2 and 3.8 (the appreciable uncertainty is due to the fact that a power law does not provide a good description of our data). Nevertheless, these rough estimates are consistent with the result of Gutin *et al.* $(\lambda = 2.7)$. (See the inset of Fig. 4 for a plot of the free energy barrier as a function of logarithm of the size of the chains.)

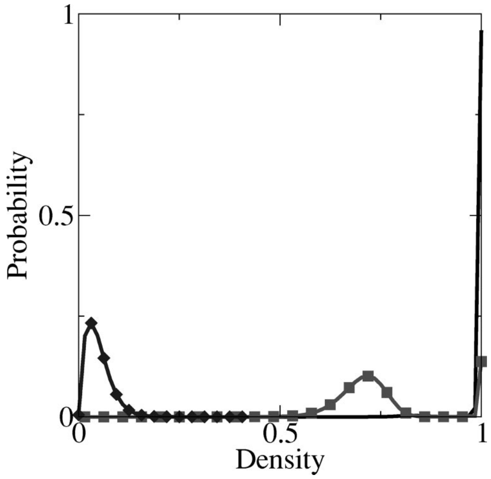

FIG. 5. Density histogram for $\epsilon/k_bT = -1.10$. The three curves correspond to $\mu/k_bT = -17.7$(diamond), $-16.5$(square), $-15.3$(solid line).

## IV. PHASE BEHAVIOR OF THE MULTIPLE-PROTEIN SYSTEM

### A. Simulations

We study the phase behavior of a system of many proteins. Since we perform simulations on a lattice, neither the constant-NPT MC, nor the Gibbs ensemble method are attractive options. Rather we simulate our system in the Grand-Canonical $(\mu VT)$ ensemble. We used parallel tempering[16,20–22] to speed up the relaxation of our systems. Usually, the parallel tempering technique simulates systems at different temperatures, exploiting the fact that systems at high temperatures may easily cross free energy barriers. Hence, by swapping the temperature between different systems, all local minima of the energy landscape are accessible. We have chosen a slightly different procedure; our systems have the same value of $\epsilon/k_bT$ but different values of $\mu/k_bT$. The idea behind this choice is that free energy barriers usually depend on the value of $\mu/k_bT$. For instance, the probability to nucleate a dense phase in a dilute phase is very high at high $\mu/k_bT$ (high supersaturation), whereas it is very low at a smaller value of $\mu/k_bT$. Thus, swapping configurations with different values of $\mu/k_bT$ also allows us to overcome free energy barriers.

We perform the simulation on a system of lattice proteins shown in Fig. 1 in a $8\times8\times8$ lattice with periodic boundaries. We stress that these proteins are very short, so much so that one should expect that this model may miss some of the features of real protein solutions. The choice of such short model proteins was based on a compromise between what is desirable and what is feasible. Tests showed that systems consisting of longer proteins got stuck in glassy states (at least on the time scales of our simulations) and this prevented the determination of the phase diagram for such molecules. While short chains may provide an oversimplified picture of proteins, we stress that our model proteins retain the two-state behavior that is one of the main aspects of the folding of the Gō model. In order to determine the phase diagram, we recorded the density histogram of each system in the parallel tempering simulation. Typical density histograms are shown in Fig. 5. The density histograms show the presence of three phases that we will respectively call ''gas,'' ''liquid,'' and ''solid.'' Phase coexistence occurs for those values of $\mu/k_bT$ where the area of each of the two peaks in the histogram are equal. An example of such a two-peaked histogram (for $\mu/k_bT = -16.5$) is shown in Fig. 5. We have used the multiple-histogram reweighting technique[23] to estimate the density histograms at intermediate values of $\mu/k_bT$. We performed simulations for a dozen different values of $\mu/k_bT$ and for six values of $\epsilon/k_bT$.

We stress that the above scheme to determine phase-coexistence works for the liquid and vapor phase, but not for the solid phase. We therefore determine the liquid–gas coexistence from the density histogram and use analytical estimates of the free energy of the solid to estimate the freezing curve.

### B. Phase diagram

To locate the coexistence between the solid and the (dilute) vapor phase, we estimate the grand partition function of the gas and of the solid analytically. The conditions of coexistence are given by the equality of the pressure, of the chemical potential $\mu/k_bT$, and of $\epsilon/k_bT$.

In our simulations, we found several possible morphologies of fully ordered solids. The structures of these three solids are represented in Fig. 6. In all three structures, each ''amino acid'' is bound to a maximum number of neighbors (4 neighbors for amino-acids in chains, and 5 for both ends).

The first solid Fig. 6(a) consists of proteins in their native structure. One can however note that the Gō model is peculiar since the native state of the protein is in fact degenerate; both the shape given in Fig. 1 and its mirror image are allowed. This is not the case in real proteins because of the chirality of the alpha carbon of the peptide-chain. However, this peculiarity of the Gō model is exploited in the solid of Fig. 6(a). We should stress here that this solid is not specific to the protein we have chosen; every protein with a compact native state can form this kind of solid within the scope of

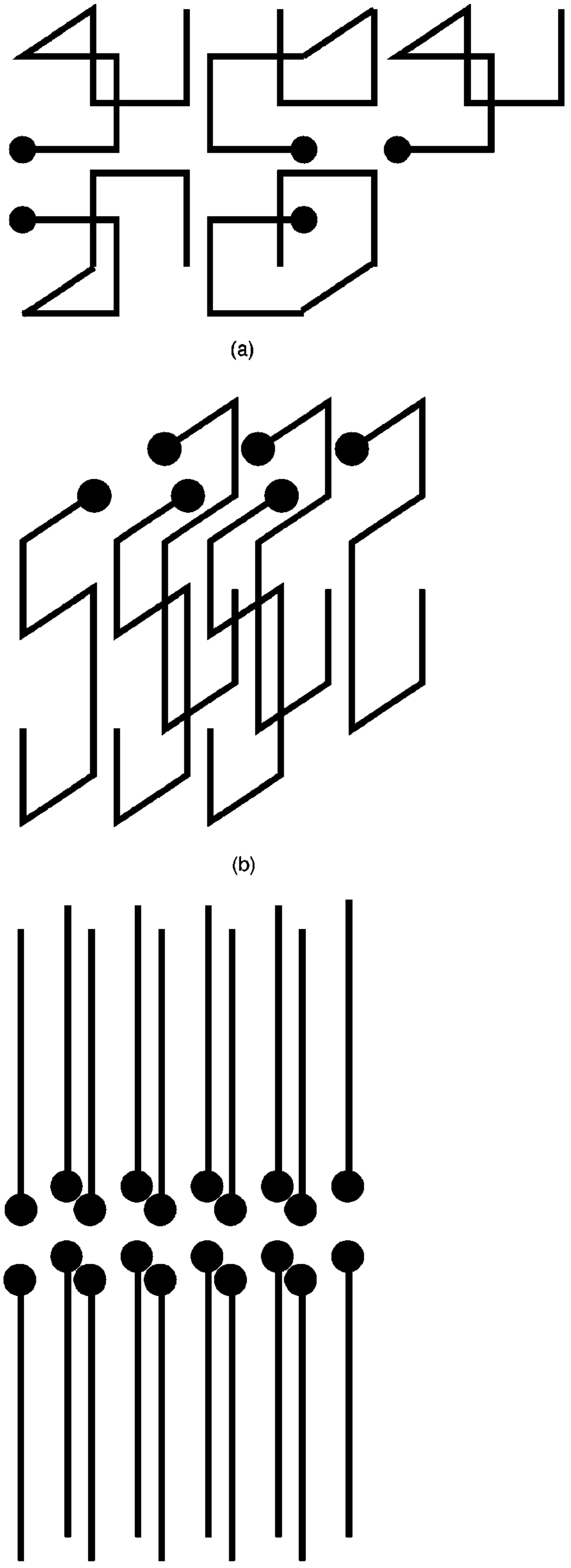

FIG. 6. Morphology of three possible crystal structures. A sphere denotes the first monomer of the chain.

the Gō model. In this solid, the number of intramolecular bonds is maximum. In the following, we call this solid the "native solid."

The second solid Fig. 6(b) is made of proteins in an "S" conformation. This solid is more likely to be specific to the protein we have chosen. In the following, we will call this solid the "S" solid.

The third solid Fig. 6(c) is made of fully stretched proteins. In this solid, the number of intermolecular bonds is at a maximum, whereas there are no intramolecular bonds. This solid is also not specific to our choice of protein. In the following, we will call this solid the "stretched solid."

To compare these three solids, we calculate their free energies. Let $N_{\mathrm{intra}}$ be the number of intramolecular bonds per chain and $N_{\mathrm{inter}}$ the number of intermolecular bonds that one chain can create with its neighbors. For the native solid, $N_{\mathrm{intra}}=5$ and $N_{\mathrm{inter}}=24$, for the S solid, $N_{\mathrm{intra}}=3$ and $N_{\mathrm{inter}}=28$ and for the stretched solid $N_{\mathrm{intra}}=0$ and $N_{\mathrm{inter}}=34$. The energy $U_0$ of a perfect crystal of volume $V=Nl_{\mathrm{seq}}$ is

$$U_0=N\left(\frac{N_{\mathrm{inter}}}{2}+N_{\mathrm{intra}}\right)\epsilon, \tag{7}$$

where $N$ is the number of proteins of length $l_{\mathrm{seq}}$ in the considered piece of the perfect crystal. One can easily see that the three solids mentioned above have exactly the same internal energy $U_0=17N\epsilon$; as already mentioned, each amino acid is bound to a maximum number of neighbors. The Grand partition function $\Xi$ of a crystal with $N$ proteins could be written

$$\Xi_{\mathrm{solid}}=e^{-\beta U_0+\beta\mu N}[1+\phi_1+\phi_2+\phi_3\cdots]. \tag{8}$$

The terms $\phi_i$ refers to the crystal with $i$ vacancies. One can then calculate each term,

$$\Phi_1=Ne^{\beta[(N_{\mathrm{intra}}+N_{\mathrm{inter}})\epsilon-\mu]}=N\zeta, \tag{9}$$

where

$$\zeta=e^{\beta[(N_{\mathrm{intra}}+N_{\mathrm{inter}})\epsilon-\mu]}. \tag{10}$$

Indeed, each time a vacancy is produced, $N_{\mathrm{intra}}+N_{\mathrm{inter}}$ bonds are broken, and there are $N$ different ways of leaving a protein. We have assumed that all proteins stay in their native conformation in the solid. We then underevaluate the number of conformation and thus the entropy. Nevertheless, these terms should not have a significant contribution, since, as we will see later, the main contribution in the free energy comes from the ground state.

The term $\Phi_2$ must involve both cases when the two vacancies are not neighbors and when they are. To simplify the calculation, we will assume that the second contribution is negligible. $\Phi_2$ then takes the simple form,

$$\Phi_2=\frac{N(N-1)}{2!}\zeta^2. \tag{11}$$

And then, from this assumption, we easily find that the term $\Phi_k$ is

$$\Phi_k=\frac{N!}{k!(N-k)!}\zeta^k. \tag{12}$$

We can then calculate easily the grand partition function from Eqs. (8) and (12),

$$\Xi_{\text{solid}}=e^{-\beta U_0+\beta\mu N}\sum_{k=0}^{N}\Phi_k \tag{13}$$

$$=e^{-\beta U_0+\beta\mu N}(1+\zeta)^N. \tag{14}$$

The grand potential and the pressure in the solid are

$$J_{\text{solid}}=-\frac{1}{\beta}\ln\Xi_{\text{solid}}, \tag{15}$$

$$P_{\text{solid}}=-\frac{J}{V} \tag{16}$$

$$=\frac{1}{\beta l_{\text{seq}}}\left[\beta\mu-\beta\left(\frac{N_{\text{inter}}}{2}+N_{\text{intra}}\right)\epsilon\right. \tag{17}$$

$$\left.+\ln(1+\zeta)\right]. \tag{18}$$

Thus far, we have neglected the fact that in that changing the orientation of one molecule without removing it, is also an excited state; it involves breaking the intermolecular bonds[31] but not the intramolecular ones. Calling $N_{\text{rot}}$ the number of possible orientations that break the intermolecular bonds of one protein in the lattice ($N_{\text{rot}}=5$ for the native solid, $N_{\text{rot}}=3$ for the S solid, and $N_{\text{rot}}=1$ for the stretched solid), the grand partition function then becomes

$$\Xi_{\text{solid}}=e^{-\beta U_0+\beta\mu N}\sum_{k=0}^{N}C_N^k\zeta^k\left[\sum_{j=0}^{N}C_{N-k}^j N_{\text{rot}}^j e^{j\beta N_{\text{inter}}\epsilon}\right], \tag{19}$$

where we have used the notation $C_N^j=N!/j!(N-j)!$,

$$\Xi_{\text{solid}}=e^{-\beta U_0+\beta\mu N}\sum_{k=0}^{N}C_N^k\zeta^k(1+N_{\text{rot}}e^{\beta N_{\text{inter}}\epsilon})^{N-k} \tag{20}$$

$$=e^{-\beta U_0+\beta\mu N}[1+N_{\text{rot}}e^{\beta N_{\text{inter}}\epsilon}+\zeta]^N. \tag{21}$$

This is exactly the same form as before, provided that we replace

$$(1+\zeta)\rightarrow(1+N_{\text{rot}}e^{\beta N_{\text{inter}}\epsilon}+\zeta), \tag{22}$$

so that

$$J_{\text{solid}}=-\frac{N}{\beta}[\beta\mu-\beta U_0+\ln(1+N_{\text{rot}}e^{\beta N_{\text{inter}}\epsilon}+\zeta)], \tag{23}$$

$$P_{\text{solid}}=\frac{1}{\beta l_{\text{seq}}}[\beta\mu-\beta U_0+\ln(1+N_{\text{rot}}e^{\beta N_{\text{inter}}\epsilon}+\zeta)]. \tag{24}$$

From Eqs. (24) and (10), we can easily check that the main contribution of the free energy comes from the internal energy of the ground state. The logarithm term is negligible compared to the energy term.

We now calculate the grand partition function and the pressure of the dilute gas. We assume a perfect gas of protein of length $l_{\text{seq}}$ in a volume $V=Nl_{\text{seq}}$, where $N$ is the maximum number of proteins in the volume. Provided that proteins do not interact in the dilute gas phase, their conformational partition function $Z_{\text{conform}}(\beta\epsilon)$ is the same as for an isolated chain.

The grand partition function $\Xi_{\text{gas}}$ is

$$\Xi_{\text{gas}}=\sum_{n=0}^{N}\frac{N!}{n!(N-n)!}(l_{\text{seq}}Z_{\text{conform}}(\beta\epsilon)e^{\beta\mu})^n \tag{25}$$

$$=(1+l_{\text{seq}}Z_{\text{conform}}(\beta\epsilon)e^{\beta\mu})^N. \tag{26}$$

In Eq. (26), $n$ is the number of proteins in the system.

From Eq. (27), we deduce the grand potential $J$ and the pressure of the gas,

$$J_{\text{gas}}=-\frac{1}{\beta}\ln\Xi_{\text{gas}} \tag{27}$$

$$=-\frac{N}{\beta}\ln(1+l_{\text{seq}}Z_{\text{conform}}(\beta\epsilon)e^{\beta\mu}), \tag{28}$$

$$P_{\text{gas}}=-\frac{J_{\text{gas}}}{V}=\frac{1}{\beta Nl_{\text{seq}}}\ln\Xi_{\text{gas}} \tag{29}$$

$$=\frac{1}{l_{\text{seq}}\beta}\ln(1+l_{\text{seq}}Z_{\text{conform}}(\beta\epsilon)e^{\beta\mu}). \tag{30}$$

From Eqs. (25) and (31), we deduce the criteria of phase coexistence,

$$P_{\text{solid}}(\beta\epsilon,\beta\mu)=P_{\text{gas}}(\beta\epsilon,\beta\mu), \tag{31}$$

which, for a given value of $\beta\epsilon$ permits us to find the value of $\beta\mu$. One can then deduce the values of the densities of each phase from the following equations (established from the partition functions)

$$d_{\text{solid}}=\frac{1}{1+N_{\text{rot}}e^{\beta N_{\text{inter}}\epsilon}+e^{\beta(N_{\text{intra}}+N_{\text{inter}})\epsilon-\beta\mu}}, \tag{32}$$

$$d_{\text{gas}}=\frac{l_{\text{seq}}Z_{\text{conform}}(\beta\epsilon)e^{\beta\mu}}{1+l_{\text{seq}}Z_{\text{conform}}(\beta\epsilon)e^{\beta\mu}}. \tag{33}$$

We calculate $Z_{\text{conform}}(\beta\epsilon)$ by the exhaustive computation of all conformations of the chains.

The full phase diagram is then shown in Figs. 7(a) and 7(b), where the solid phase is the native solid. The phase diagram shows that the gas–liquid phase transition is metastable. During our simulations, we found direct evidence that the liquid phase is metastable with respect to the crystal; it appears to be an ‘‘intermediate’’ state between the vapor and the solid. Of course, our estimate for the gas–solid coexistence will be incorrect at high vapor densities. While this will change the high-temperature solid–vapor transition curve, it will not affect our conclusions regarding the metastability of the gas–liquid phase transition. In fact, this phase diagram is in qualitative agreement with the protein phase diagram found both in experiments and in theoretical studies.[24]

## V. DISCUSSION AND CONCLUSION

We have presented the calculation of the phase diagram of both isolated chains and multiproteins systems in the scope of the Gō model. The study of isolated chains shows a two states behavior: the coil and native conformations. We have shown that, as the chain lengths increases, the transition between these two conformations tend to a first order transition.

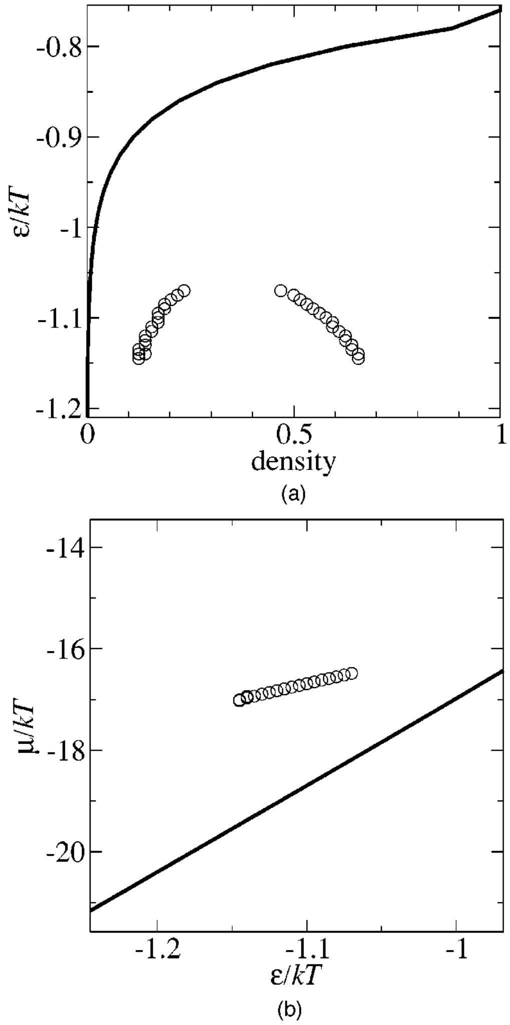

FIG. 7. Phase diagram. (a) In the $(\epsilon/k_bT,\text{density})$ coordinates. (b) In the $(\mu/k_bT,\epsilon/k_bT)$ coordinates. The gas–liquid transition (circle) is metastable. The gas–solid coexistence is plotted with a solid line.

The calculation of the phase behavior of a three-dimensional system consisting of many, proteinlike chain molecules, is extremely demanding, even for the present, highly simplified, model. This may well explain why, to our knowledge, no such simulation has been reported thus far (see, however, Bratko *et al.*[25] and Smith *et al.*,[26] where they, respectively, simulate 6 and 4 proteins but using a bidimensional lattice[8,27,28] or a intermediate resolution model). Our multiprotein model exhibits three phases: vapor, liquid, and crystal. However, the liquid–vapor transition occurs at a temperature below freezing. In the gas phase, the proteins hardly interact, although dimers are sometimes observed, depending on the density.

The liquid phase is a disordered structure consisting of partially folded proteins. To analyze the molecular conformations in this phase, we compared the probabilities to find $i$ intramolecular bonds in an isolated chain and in the liquid phase. We found that proteins in the liquid phase are slightly more compact than isolated chains. For instance, at $\epsilon/k_bT=-1.1$, the probability that an isolated protein has no native bond is 24.4%, whereas it is 17.4% in the liquid at $\mu/k_bT=-16.2$. The probability that an isolated chain in its native state is 6.3%, whereas it is 11.6% in the liquid. Thus, event though the notion of partially folded chain is, in the present case, ill defined, as our chains are very short, the average number of intramolecular bonds per chain is higher in the liquid phase than for the isolated proteins. We can therefore conclude that the liquid phase, is composed of partially folded proteins. This observation is in agreement with previous observations on a two-dimensional model system.[29] Moreover, as the transition between the coil and the native state of isolated proteins occurs for $\epsilon/k_bT=-1.73$, the liquid phase stabilizes some conformations partially folded that would not be stable for an isolated chain. The effect of density on folding is even more striking in the solid. We find that crystallization drives proteins either to their native conformation, or to another, very specific, conformation (in our case, S-shaped or linear). Within our model, the three solid structures have almost the same free energy and we have actually observed in the simulations that a spontaneously formed solid is a mixture of the "native" solid and of the "S" solid. We have not observed spontaneous formation of "extended chain" crystals. The absence of extended chain crystals could either be due to kinetic factors (as in the case of homopolymers[30]), or to finite size effects. However, in larger system we also did not observe the spontaneous formation of extended-chain crystals. This suggests that, even for very short chains, kinetic effects are important in determining the crystal structure.

Thus far we have assumed that the strength of intermolecular interactions is equal to that of intramolecular interactions. This is clearly an oversimplification: one would expect that the interaction between hydrophilic surface groups on different molecules would be rather weak. The more so as, in real proteins crystals,[17] some water still separates the protein surfaces, whereas the hydrophobic effect leads to an expulsion of water from the protein core. On the other hand, one might expect to observe strong intermolecular interactions between the hydrophobic residues in two unfolded proteins. To explore the effect of a change in the relative strength of inter and intramolecular interactions, we have performed some preliminary simulations with modified intermolecular interactions. The main effect of these changes appears to be an overall vertical shift of the computed phase diagram; increasing the strength of the intermolecular interactions stabilizes the denser phases.

It is clear that the $\overline{\text{Go}}$ model represents an oversimplified picture of a system of proteins. Nevertheless, it allows us to reproduce some of the qualitative features of real protein system: two-state behavior for isolated chains and a phase diagram that contains a metastable gas–liquid coexistence curve. One serious drawback of the simple $\overline{\text{Go}}$ model is that the molecules are nonchiral. This results in unrealistic crystal structures that favor the native state. In order to study the possible aggregation of unfolded model proteins, it would be necessary to employ a model that does not possess such a spurious reflection symmetry.

## ACKNOWLEDGMENTS

One of us (N.C.) is grateful to P.R. Ten Wolde for useful discussions. This research has been supported by a Marie

Curie Fellowship of the European Community program ‘‘Improving Human Research Potential and the Socio-Economic Knowledge Base’’ under Contract No. HPMF-CT-2001-01212. Disclaimer: the authors are solely responsible for information communicated and the European Commission is not responsible for any views or result expressed. The work of the FOM Institute is part of the research program of FOM and is made possible by financial support from the Netherlands organization for Scientific Research (NWO).

[1] R. C. Moore and D. W. Melton, Mol. Hum. Reprod. **3**, 529 (1997).
[2] A. Slepoy, R. R. P. Singh, F. Pázmándi, R. V. Kulkarni, and D. L. Cox, Phys. Rev. Lett. **87**, 058101/1 (2001).
[3] L. K. Simmons, P. C. May, K. J. Tomoselli *et al.*, Mol. Pharmacol. **45**, 373 (1994).
[4] D. J. Selkoe, J. NIH Res. **7**, 57 (1995).
[5] J. Clark and J. Steele, Proc. Natl. Acad. Sci. U.S.A. **89**, 1720 (1992).
[6] H. R. Costantino, R. Langer, and A. M. Klibanov, Biotechnology **13**, 493 (1995).
[7] K. M. Personn and V. Gekas, Process Biochem. **29**, 89 (1994).
[8] P. Gupta, C. K. Hall, and A. C. Voegler, Protein Sci. **7**, 2642 (1998).
[9] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, 3rd ed. (Garland, New York, 1994).
[10] N. Go, J. Stat. Phys. **30**, 413 (1983).
[11] S. Takada, Proc. Natl. Acad. Sci. U.S.A. **96**, 11698 (1999).
[12] H. S. Chan and K. A. Dill, Proteins **30**, 2 (1998).
[13] D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, New York, 2000).
[14] P. H. Verdier, J. Chem. Phys. **59**, 6119 (1973).
[15] A. Sali, E. I. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[16] D. Frenkel and B. Smit, *Understanding Molecular Simulation*, 2nd ed. (Academic, London, 2002).
[17] C. Branden and J. Tooze, *Introduction to Protein Structure*, 2nd ed. (Garland, New York, 1998).
[18] R. Guerois and L. Serrano, Curr. Opin. Struct. Biol. **11**, 101 (2001).
[19] A. Gutin, V. Abkevich, and E. I. Shakhnovich, Phys. Rev. Lett. **77**, 5433 (1996).
[20] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov, J. Chem. Phys. **96**, 1776 (1992).
[21] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).
[22] C. J. Geyer and E. A. Thompson, J. Am. Stat. Assoc. **90**, 909 (1995).
[23] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **61**, 2635 (1988).
[24] N. Asherie, A. Lomakin, and G. B. Benedek, Phys. Rev. Lett. **77**, 4832 (1996).
[25] D. Bratko and H. W. Blanch, J. Chem. Phys. **114**, 561 (2001).
[26] A. V. Smith and C. K. Hall, J. Mol. Biol. **312**, 187 (2001).
[27] R. I. Dima and D. Thirumalai, Protein Sci. **11**, 1036 (2001).
[28] P. M. Harrison, H. S. Chan, S. B. Prusiner, and F. E. Cohen, Protein Sci. **10**, 819 (2001).
[29] P. Gupta, C. K. Hall, and A. Voegler, Fluid Phase Equilib. **158–160**, 87 (1999).
[30] J. I. Lauritzen and J. D. Hoffman, J. Res. Natl. Bur. Stand., Sect. A **64A**, 73 (1960).
[31] We will assume here that all of the intermolecular interactions are broken.