Folding simulations of a three-dimensional protein model with a nonspecific hydrophobic energy function

Leandro G. Garcia, Werner L. Treptow, and Antônio F. Pereira de Araújo
Departamento de Biologia Celular and International Center of Condensed Matter Physics, Universidade de Brasília, Brasília-DF 70910-900, Brazil

(Received 31 January 2001; published 21 June 2001)

We show that a nonspecific hydrophobic energy function can produce proteinlike folding behavior of a three-dimensional protein model of 40 monomers in the cubic lattice when the native conformation is chosen judiciously. We confirm that monomer inside/outside segregation is a powerful criterion for the selection of appropriate structures, an idea that was recently proposed with basis on a general theoretical analysis and simulations of much simpler two-dimensional models.

DOI: 10.1103/PhysRevE.64.011912 PACS number(s): 87.15.Cc, 87.10.+e

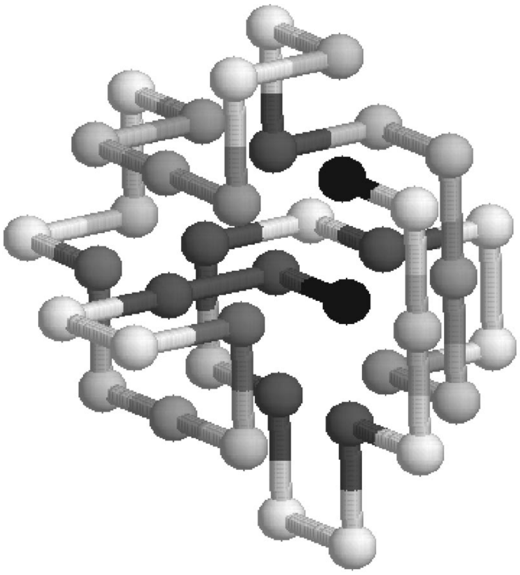
How do proteins fold? This question has been one of the most challenging problems of molecular biophysics during the last decades of the 20th century. A general answer can be provided in terms of a rugged, funnel-shaped, energy surface that should be able to rapidly guide the ensemble of unfolded conformations towards the native structure [1–4]. It is much less understood, however, how this appropriate surface arises during the folding process. Hydrophobicity must necessarily be involved since it is known to be the most important factor determining protein stability [5], but can the intrinsically nonspecific hydrophobic effect, by itself, encode unique native structures inside astronomically large conformational spaces? Results from previous studies on lattice models, with monomers intended to mimic polar and apolar amino acids, have been controversial. Although complete enumeration of short chains indicates the possibility of proteinlike thermodynamics, attempts to design longer sequences to fold to maximally compact conformations tend to fail unless pair-specific segregation terms are added to the energy function [6–9].

It is clear that hydrophobic amino acids of a protein in water will tend to hide from the solvent while the reverse will be true for hydrophilic amino acids. A general and simple lattice implementation of this idea is a conformational energy function in the form of the negative scalar product in N-dimensional space between the sequence, represented by the vector $\vec{h}=\{h_1,\ldots,h_N\}$, and the structure, represented by the vector $\vec{c}=\{c_1,\ldots,c_N\}$, where h_i is the hydrophobicity (positive or negative) of monomer i and c_i is the number of contacts it makes [8,10–13],

$$E(\{h_i\},\{c_i\})=-\sum_{i=1}^{N} h_i c_i=-\vec{h}\cdot\vec{c}. \tag{1}$$

A recent analysis [12] based on the stability criterion for proteinlike folding behavior [14–17,7] suggested that, although nonspecific, in the sense that the contribution of a monomer participating in a contact is independent of its contact partner, this function could be successfully used in folding simulations. Possible native structures were predicted to be not arbitrary, however, since they should have large values of σ, the standard deviation of the number of contacts made by each of their monomers. Native structures, therefore, should have their monomers sufficiently segregated between buried positions, making the maximal number of contacts, and exposed positions, making no contacts at all. Interestingly, maximally compact conformations in the square lattice were found to be not adequate since many of their monomers have an intermediate degree of exposure, resulting in low values of σ [12]. Although successfully tested with chains of 24 monomers in the two-dimensional square lattice, a more general implementation of this idea in three dimensions was lacking until now, mainly because of the difficulty to generate sufficiently segregated three-dimensional conformations.

In the present study, a segregated conformation of 40 monomers in the three-dimensional cubic lattice was generated by a specific computer program (Fig. 1). The searching algorithm consisted of a standard Monte Carlo simulation



54400312130102122201041121200404140040105

FIG. 1. Native conformation used in the present study. The number below the structure represents its contact vector $\vec{c}$, i.e., each digit represents the number of contacts in the native conformation made by each monomer along the sequence. These numbers are also indicated on the structure itself by different shades of gray, ranging from white (0 contacts) to black (5 contacts).

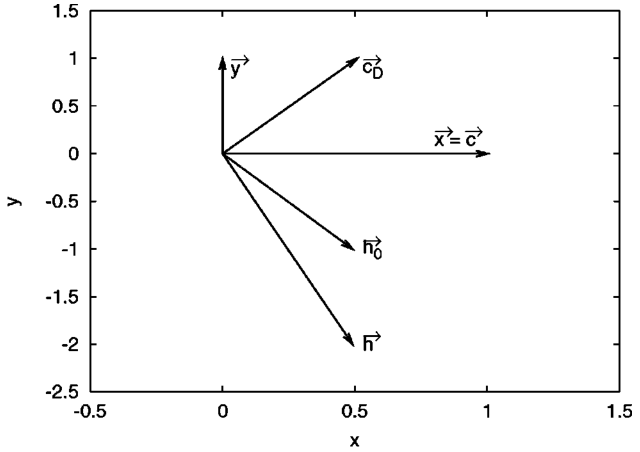


FIG. 2. Vectorial representation of the sequence design procedure in the sequence-structure diagram. The plane represents the two-dimensional subspace defined by the contact vector $\vec{c}$ and its diagonal projection $\vec{c}_D=\{\overline{c},\ldots,\overline{c}\}$, where $\overline{c}=1.65$ is the average over the components of $\vec{c}$. Vectors $\vec{x}=\vec{c}$ and $\vec{y}=\vec{c}_D-K\vec{c}$, where $K=(\vec{c}.\vec{c}_D/|\vec{c}|^2)=0.509$ is a conformation-dependent constant, are conveniently taken as an orthogonal basis in this plane. Units represented along each axis are different, since $\vec{x}$ and $\vec{y}$ have different lengths. A sequence perpendicular to the diagonal is obtained from $\vec{h}_0=\vec{c}-\vec{c}_D=(1-K)\vec{x}-\vec{y}$. The sequence used in the present study was slightly rotated away from this direction by doubling its $\vec{y}$ component, $\vec{h}=(1-K)\vec{x}-2\vec{y}$. The resulting hydrophobicities of each monomer are $h(0)=-3.3$, $h(1)=-1.791$, $h(2)=-0.282$, $h(3)=1.227$, $h(4)=2.736$, and $h(5)=4.245$, where $h(c)$ stands for the hydrophobicity of monomers making c contacts in the native structure.

with the Metropolis criterion [18] as previously implemented [19,12] but with an energy function depending explicitly on σ. The temperature was lowered very slowly in an attempt to find deep minima in the resulting energy surface, which should correspond to conformations with the desired property. Although not completely segregated, the obtained σ of 1.62 is significantly higher than 0.85, the corresponding value for maximally compact conformations of 36 monomers, like the one used in previous studies with more specific pairwise contact energies [19]. If a conformation could be completely segregated between the minimal of 0 and the maximal of 4 contacts (except for the chain ends) its σ would be 2, but such conformation cannot exist in the cubic lattice due to topological constraints and it is difficult to predict how far below is 1.62 from the topologically possible limit. Note that although having no ''cavities,'' the segregated structure has a total of only 33 contacts resulting in an average of 0.825 contacts per monomer, much smaller than 1.111, the corresponding average for the maximally compact conformation.

The hydrophobic sequence intended to fold to the generated structure was obtained as a vector $\vec{h}$ in the plane defined by the conformation contact vector $\vec{c}$ and the main diagonal [12,13]. The angle between $\vec{h}$ and the diagonal was not 90° but slightly larger, since this small ''rotation'' was recently found to improve folding behavior for two-dimensional models [13]. The sequence design procedure can be visualized in a two-dimensional sequence-structure diagram (Fig. 2). Sequence rotation emphasizes the importance of some effectively repulsive non-native interactions and its effect on folding behavior is probably related to the recently observed smoothening effect caused by such repulsive terms on the energy surfaces of simple models [20]. Representative folding trajectories of 2.5 billion time steps recorded every 100 000 steps (1 time step = 40 move attempts) of the designed sequence, using the hydrophobic energy function [Eq. (1)] and beginning from random initial conformations, are shown in Fig. 3. Folding is very cooperative to the correct native structure and no other conformation with equal or lower energy was ever found. Kinetics around the folding temperature appear to be at least two orders of magnitude slower when compared to previous three-dimensional models with pairwise contact interactions [19], suggesting a

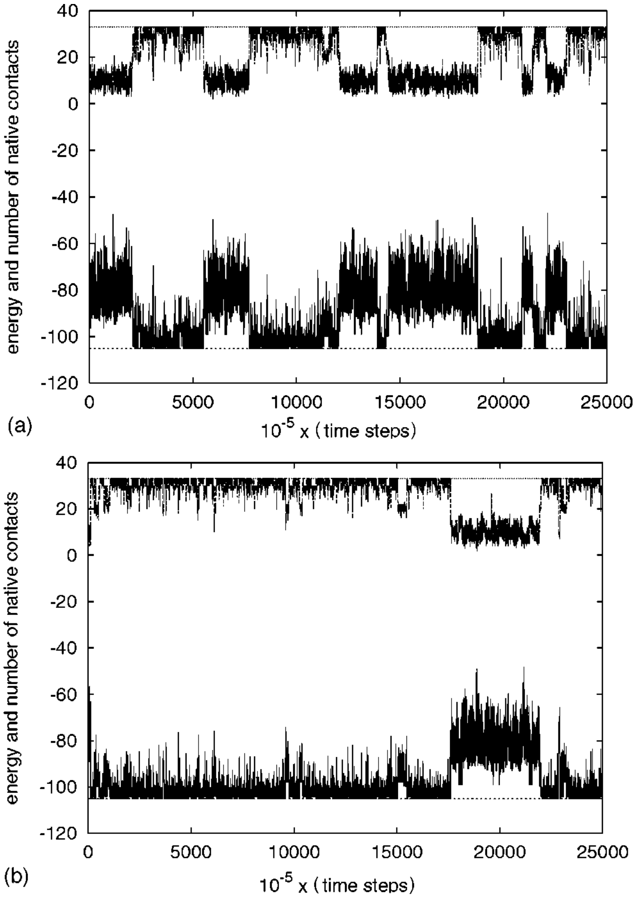


FIG. 3. Representative trajectories at $T=1.26=T_f$ (a) and $T=1.15<T_f$ (b), obtained by standard Monte Carlo simulations for the designed sequence with the hydrophobic energy function. Energy and number of native contacts are plotted against the number of time steps (1 time step = 40 move attempts). The energy (-105.1) and number of contacts (33) of the native structure are shown as horizontal lines. Temperature is always represented in the present study in units of energy (i.e., Boltzmann contant is taken to be unity) and the energy scale is determined by the numerical values of the hydrophobicities along the sequence.

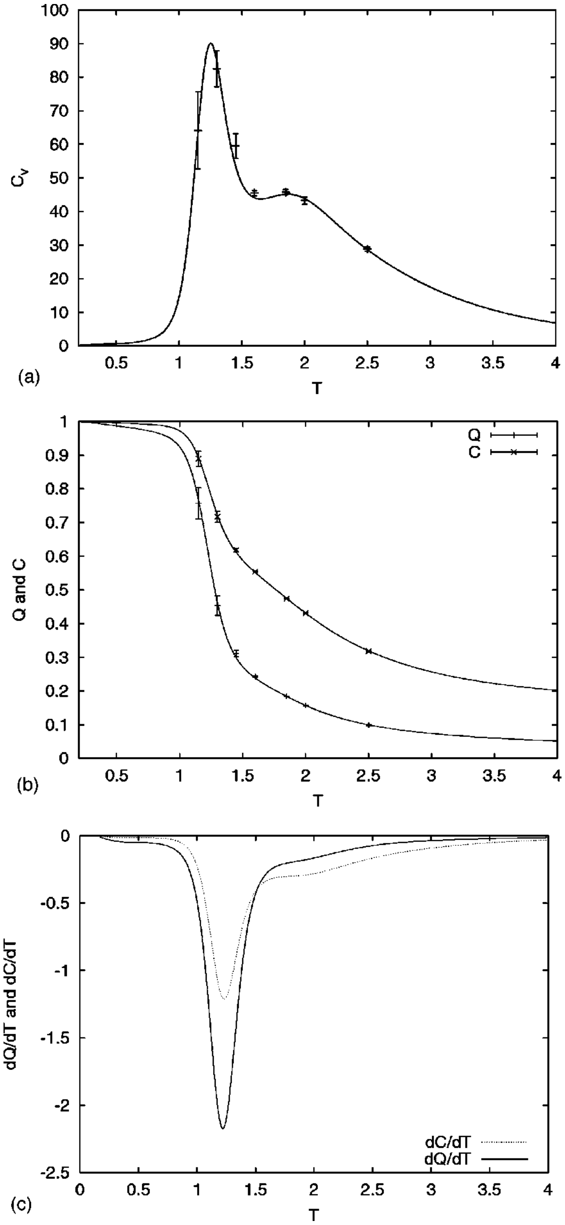


FIG. 4. Heat capacity (a), average fraction of native contacts, Q, and average fraction of all contacts, C, (b), and the derivatives of the average contact fractions (c) as a function of temperature. Curves shown in (a) and (b) were obtained by the multiple histogram technique from long simulations of up to 10 billion time steps (1 time step corresponds to $N=40$ move attempts) run at $T=1.30$, $T=1.45$, $T=1.60$, $T=1.85$, $T=2.0$, and $T=2.5$. Points and error bars shown in (a) and (b) represent the average and standard deviation over independent simulations at these temperatures and also for four simulations of 5 billion time steps run at $T=1.15$ that were not used in the multiple histogram procedure. Good agreement between points and curves, even for $T=1.15$, confirms that the extrapolation was reasonably accurate. Curves shown in (c) were obtained directly from the curves shown in (b). Q and C are adimensional and, since temperature has units of energy, the heat capacity $C_v=dE/dT$ is also adimensional while dQ/dT and dC/dT have units of inverse energy. The energy scale is determined by the numerical values of the hydrophobicities along the sequence.

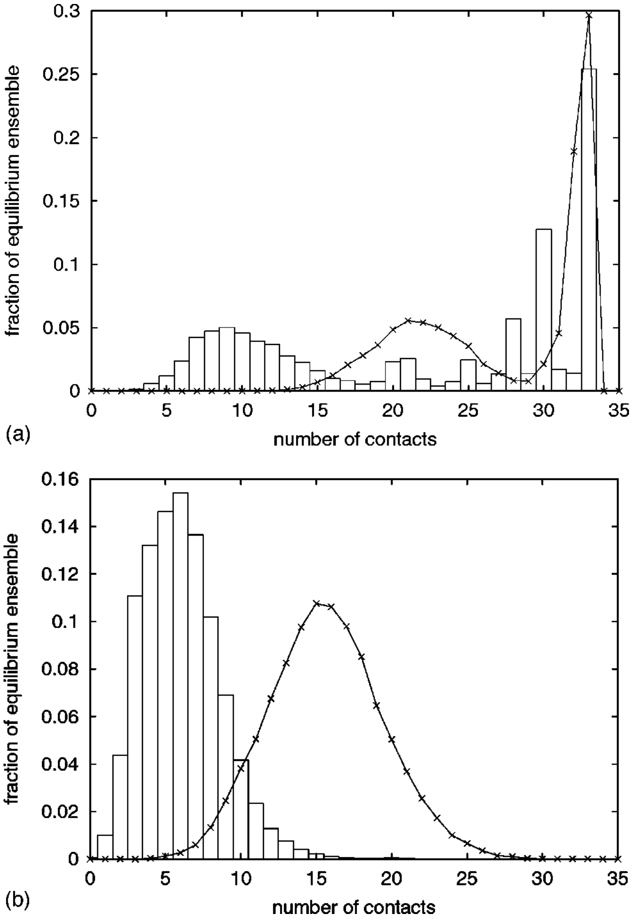


FIG. 5. Equilibrium distributions of the number of native contacts (bars) and the number of all contacts (lines) at $T=1.26$, the folding transition temperature (a), and at $T=1.85$, around the middle of the discompactization transition (b). The number of native contacts and all contacts are simply Q and C, respectively, multiplied by 33, the number of contacts in the native conformation.

higher free-energy barrier for folding. Despite the requirement of longer simulations, proteinlike thermodynamics are clearly reproduced. At high temperatures the chain stays most of the time in the unfolded state (not shown). At $T=T_f=1.26$, the folding transition temperature, both the unfolded and native states are significantly populated [Fig. 3(a)]. At $T=1.15$, below the transition temperature, the chain folds and then remains most of the time in the folded state [Fig. 3(b)], indicating the native structure is kinetically accessible at a temperature where it is thermodynamically stable.

The temperature dependence of relevant thermodynamic averages were obtained by the multiple histogram technique [21,22] from several long simulations run at different temperatures. The heat capacity displays a sharp peak at $T=1.26$, which is used to define the folding transition temperature T_f and a shoulder at $T\approx1.85$ [Fig. 4(a)]. The average fraction of all contacts, C, a measure of chain compactness, displays the same behavior of the heat capacity with a sharp variation around T_f, followed by a more gradual variation at higher temperatures [Fig. 4(b)], implying a sharp peak followed by a shoulder in its derivative [Fig. 4(c)]. For the average fraction of native contacts, Q, a measure of the amount of native structure, a single sharp transition is apparent [Fig. 4(b)] resulting in a single sharp peak in its derivative at $T = T_{\mathrm{nc}} \approx T_f$ [Fig. 4(c)]. It appears, therefore, that the native structure is disrupted in a single cooperative transition, while the shoulder on the heat capacity curve results from further discompactization of the unfolded state.

Reasonably bimodal equilibrium distributions of Q and C at $T = 1.26$ confirm that the folding transition is first order (''two-state'') [Fig. 5(a)] while their unimodal distributions at $T = 1.85$ indicate that the discompactization transition is second order [Fig. 5(b)]. Very similar behavior has been previously reported for different models of protein folding [23]. The discompactization transition has even been found in simple homopolymer models [24] but is not detected in heat capacity curves of real single-domain proteins [25]. This discrepancy might indicate that our model is more energetically frustrated than real proteins due to the small number of monomer types, since nonspecific hydrophobic interactions keep the unfolded state in a collapsed form if the temperature is not too high. It might also be relevant, however, that hydrophobic interactions themselves are intrinsically dependent on temperature [25] and such dependence is not being considered in the present study.

An adimensional cooperativity index for the folding transition can be computed as proposed by Klimov and Thirumalai [22] from the derivative of Q by the expression

$$\Omega_c = T_{\mathrm{nc}}^2 \frac{\max\left[-\dfrac{dQ}{dT}\right]}{\Delta T}, \tag{2}$$

where T_{nc} is the temperature at which the peak occurs and ΔT is the width of the peak at its half-maximal height. We have obtained the value $\Omega_c \approx 11$ for the present model, which is much higher than values obtained for the hydrophobic model in two dimensions, where it can range from 2 to 6 [13], and also for the three-dimensional model with side chains of Klimov and Thirumalai, where it is around 5 for the most cooperative sequences [22].

It should be emphasized that the structural segregation criterion for the selection of native conformations was derived from a theoretical analysis completely independent of lattice geometry. In simple terms, a segregated structure should be appropriate because hydrophobic monomers in buried positions would result in a larger decrease in its energy than in the average energy of the unfolded state while hydrophilic monomers in exposed positions would not increase its energy as much as the energy of the unfolded state. The conclusion that maximally compact conformations are not sufficiently segregated, however, results exclusively from our observations in square and cubic lattices and might, therefore, be not as general. It is not impossible, *a priori*, that for a different lattice or for some off-lattice models, maximally segregated conformations would also be, coincidentally, maximally compact. In either case, this study points out the possibility that the amount of compactness observed in globular proteins is an indirect consequence of structural segregation and not of fundamental physical significance in itself.

Taken together, our results suggest that a nonspecific hydrophobic effect can, in principle, determine the native structure of protein molecules. They support the basic premise of simpler (and limited to short chains) exact models (reviewed in [26]) that unique protein conformations do not arise from specific interactions between their monomers but from the specific pattern of hydrophobicities along the polymer chain. The possibility of folding long three-dimensional chains and the general procedure for obtaining ideal sequences for appropriate structures have broad potential implications, ranging from the use of more realistic models in the interpretation of experimental results to the design of real protein sequences. Further studies will be required to explore these possibilities.

[1] P. Leopold, M. Montal, and J. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).

[2] J. Bryngelson, J. Onuchic, N. Socci, and P. Wolynes, Proteins: Struct., Funct., Genet. **21**, 167 (1995).

[3] P. Wolynes, J. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995).

[4] K. Dill and H. Chan, Nat. Struct. Biol. **4**, 10 (1997).

[5] K. Dill, Biochemistry **29**, 7133 (1990).

[6] K. Yue *et al.*, Proc. Natl. Acad. Sci. U.S.A. **92**, 325 (1995).

[7] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).

[8] M. Skorobogatiy, H. Guo, and M. J. Zuckermann, Macromolecules **30**, 3403 (1997).

[9] R. Melin, H. Li, N. S. Wingreen, and C. Tang, J. Chem. Phys. **110**, 1252 (1999).

[10] H. Li, R. Helling, C. Tang, and N. S. Wingreen, Science **273**, 666 (1996).

[11] H. Li, C. Tang, and N. S. Wingreen, Proc. Natl. Acad. Sci. U.S.A. **95**, 4987 (1998).

[12] A. F. Pereira de Araújo, Proc. Natl. Acad. Sci. U.S.A. **96**, 12482 (1999).

[13] A. F. Pereira de Araújo, J. Chem. Phys. **114**, 570 (2001).

[14] R. Goldstein, Z. Luthey-Schulten, and P. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **89**, 4918 (1992).

[15] A. Šali, E. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).

[16] E. Shakhnovich and A. Gutin, Nature (London) **93**, 2043 (1990).

[17] E. Shakhnovich and A. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195 (1993).

[18] N. Metropolis, A. Rosembluth, M. Rosembluth, and A. Teller, J. Chem. Phys. **21**, 1087 (1953).

[19] A. F. Pereira de Araújo and T. C. Pochapsky, Folding Des. **1**, 299 (1996).

[20] M. Li and M. Cieplak, Eur. Phys. J. B **14**, 787 (2000).

[21] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).

[22] D. Klimov and D. Thirumalai, Folding Des. **3**, 127 (1998).

[23] E. Shakhnovich, G. Farztdinov, A. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).

[24] Y. Zhou, C. Hall, and M. Karplus, Phys. Rev. Lett. **77**, 2822 (1996).

[25] P. L. Privalov, in *Protein Folding*, edited by T. E. Creighton (Freeman, San Franciso, 1992), Chap. 7.

[26] K. Dill *et al.*, Protein Sci. **4**, 561 (1995).