

# Folding protein models with a simple hydrophobic energy function: The fundamental importance of monomer inside/outside segregation

Antônio F. Pereira de Araújo<sup>†</sup>

Departamento de Biologia Celular and International Center of Condensed Matter Physics, Universidade de Brasília, Brasília-DF 70910-900, Brazil

Edited by Peter G. Wolynes, University of Illinois at Urbana–Champaign, Urbana, IL, and approved August 5, 1999 (received for review May 3, 1999)

The present study explores a “hydrophobic” energy function for folding simulations of the protein lattice model. The contribution of each monomer to conformational energy is the product of its “hydrophobicity” and the number of contacts it makes, i.e.,  $E(\vec{h}, \vec{c}) = -\sum_{i=1}^N c_i h_i = -(\vec{h} \cdot \vec{c})$  is the negative scalar product between two vectors in  $N$ -dimensional cartesian space:  $\vec{h} = (h_1, \dots, h_N)$ , which represents monomer hydrophobicities and is sequence-dependent; and  $\vec{c} = (c_1, \dots, c_N)$ , which represents the number of contacts made by each monomer and is conformation-dependent. A simple theoretical analysis shows that restrictions are imposed concomitantly on both sequences and native structures if the stability criterion for protein-like behavior is to be satisfied. Given a conformation with vector  $\vec{c}$ , the best sequence is a vector  $\vec{h}$  on the direction upon which the projection of  $\vec{c} - \bar{c}$  is maximal, where  $\bar{c}$  is the diagonal vector with components equal to  $\bar{c}$ , the average number of contacts per monomer in the unfolded state. Best native conformations are suggested to be not maximally compact, as assumed in many studies, but the ones with largest variance of contacts among its monomers, i.e., with monomers tending to occupy completely buried or completely exposed positions. This inside/outside segregation is reflected on an apolar/polar distribution on the corresponding sequence. Monte Carlo simulations in two dimensions corroborate this general scheme. Sequences targeted to conformations with large contact variances folded cooperatively with thermodynamics of a two-state transition. Sequences targeted to maximally compact conformations, which have lower contact variance, were either found to have degenerate ground state or to fold with much lower cooperativity.

Current theories of macromolecular folding predict that an important factor determining “protein-like” behavior is that the native structure corresponds to a sufficiently deep global minimum in the folding energy landscape (1–4). From an operational perspective, the simplest mathematical expression for this idea is the requirement of a large negative “ $Z$ -score” that, in analogy to elementary statistics, represents the difference between the energy of the native structure and the average energy of unfolded conformations in terms of the corresponding standard deviation (5–7). If sequences of  $N$  monomers are represented by an ordered set of  $N$  letters,  $\{a_i\} = \{a_1, \dots, a_N\}$ , where  $a_i$  stands for the monomer type at position  $i$  along the sequence, and conformations are represented by the set of  $N$  vectors,  $\{\vec{r}_i\} = \{\vec{r}_1, \dots, \vec{r}_N\}$ , where  $\vec{r}_i = (x_i, y_i, z_i)$  are the coordinates of monomer  $i$ , then a given sequence  $\{a_i\}^*$  will be likely to fold to a given conformation  $\{\vec{r}_i\}^*$  if  $\{a_i\}^*$  and  $\{\vec{r}_i\}^*$  correspond to a large negative  $Z$ -score:

$$Z = \frac{E(\{a_i\}^*, \{\vec{r}_i\}^*) - \bar{E}}{\sigma_E}, \quad [1]$$

where  $E(\{a_i\}^*, \{\vec{r}_i\}^*)$  is the energy of the given sequence in the given conformation and  $\bar{E}$  and  $\sigma_E$  are the average and standard deviation of the energy of the given sequence distributed among the unfolded state ensemble.

The simplifying assumption of statistical independence of energies among different conformations results in the so-called Random Energy Model (REM) (8, 9), which was originally proposed for spin glasses (10) and shown to be a reasonable approximation for protein lattice models (11–15) (for recent theoretical reviews, including the REM, see refs. 4, 16, and 17). The REM’s most important prediction is that thermodynamic properties of the unfolded state, which is dominated by high energy conformations, do not depend on sequence but only on monomer composition. The energy of the native structure, on the other hand, which is by definition the lowest energy conformation, is sequence dependent. If the energy function is defined,  $E(\{a_i\}^*, \{\vec{r}_i\}^*)$  can be computed directly, while  $\bar{E}$  and  $\sigma_E$  can be easily estimated from the composition of  $\{a_i\}^*$ . Protein-like behavior arises when  $Z$  is much larger, in absolute value, than  $Z_c$ :

$$|Z| \gg |Z_c|, \quad [2]$$

where

$$Z_c = \frac{E_c - \bar{E}}{\sigma_E} = -\sqrt{2S_0} \quad [3]$$

is the  $Z$ -score corresponding to the critical energy  $E_c$ , the energy below which thermodynamics become sequence dependent.  $Z_c$  is another sequence-independent quantity, but the crucial value of  $S_0$ , the entropy of the unfolded state, may not be easy to estimate *a priori*.

The above scheme is very general, as it does not make any assumptions about how different conformations are represented or about the form of the energy function. Many theoretical studies and lattice model simulations have used an energy function of the form

$$E(\{a_i\}, \{\vec{r}_i\}) = \sum_{j>i=1}^N B_{a_i a_j} \Delta(i, j), \quad [4]$$

where  $\Delta(i, j)$  is 1 if monomers  $i$  and  $j$  form a contact and 0 otherwise. In the lattice model,  $i$  and  $j$  are said to form a contact if  $\vec{r}_i$  and  $\vec{r}_j$  are adjacent lattice sites and  $|j - i| > 1$ .  $B_{a_i a_j}$  is taken from a symmetric  $L \times L$  matrix of energy interactions,  $B$ , containing the energetic contribution of all possible pairwise interactions.  $L$  is the number of available monomer types, or “letters,” and has ranged from 2, as it does in the HP model, to  $N$ , as it does in the independent interaction model (reviewed in ref. 17).

Once  $B$  is chosen it is possible to use Eqs. 1 and 4 to produce sequences with good  $Z$ -scores in a particular conformation. The stability criterion can be easily satisfied for arbitrarily chosen maximally compact conformations if the number of letters is large enough. Many studies have used sequences selected to fold

This paper was submitted directly (Track II) to the PNAS office.

<sup>†</sup>E-mail: aaraujo@unb.br.

to maximally compact conformations in the cubic lattice using an alphabet of 20 monomer types (reviewed in ref. 18). Sequences have been selected either by direct  $Z$ -score minimization or by energy minimization with fixed monomer composition, which is also equivalent to minimization of Eq. 1 because in this last case  $\bar{E}$  and  $\sigma_E$  are constants. Well selected sequences of up to at least 80 monomers (3) fold cooperatively in Monte Carlo simulations to the target structure starting from any randomly generated initial conformation and the target structure is thermodynamically stable at a temperature where folding is fast (18). This clear protein-like behavior indicates that Eq. 2 is indeed a valid selection criterion.

The resulting scheme permits some objections, however (see, for example, ref. 19). First, although real proteins are made up of 20 monomer types, it is well known that amino acids can be grouped in a much smaller number of classes sharing similar physical chemical properties, such as aliphatics, aromatics, polar and charged, for example, or, even more simply, hydrophobics and hydrophilics. This suggests that fundamental properties of the folding process could be encoded by a smaller, physically meaningful, alphabet. Second, the choice of the target conformation, which plays the role of the native structure in the model, appears to be completely arbitrary, as far as it is maximally compact. The model does not have anything to say about the fact that native structures of globular proteins tend to fall into a relatively small number of folds.

This study combines the general stability criterion for sequence selection with a simple and physically intuitive energy function for the protein lattice model. Energetic contributions for possible contacts are taken from an interaction matrix with elements of the form

$$B_{a_i a_j} = -(h(a_i) + h(a_j)) = -(h_i + h_j), \quad [5]$$

where  $h_i = h(a_i)$  is the “hydrophobicity” of the monomer type  $a_i$ . Hydrophobicities can be either positive or negative. The additive form of the  $B$  matrix elements shown in Eq. 5 implies that Eq. 4 can be expressed simply as (1, 5, 20)

$$E(\{a_i\}, \{\vec{r}_i\}) = E(\vec{h}, \vec{c}) = \sum_{i=1}^N -(h_i c_i) = -(\vec{h} \cdot \vec{c}), \quad [6]$$

where  $\vec{h} \cdot \vec{c}$  is the scalar product between two vectors in  $N$ -dimensional cartesian space: the hydrophobicity vector,  $\vec{h} = (h_1, \dots, h_N)$ , which represents monomer hydrophobicities and depends only on the sequence, and the contact vector,  $\vec{c} = (c_1, \dots, c_N)$ , which represents the number of contacts made by each monomer and is therefore dependent on the conformation. Note that the function is “unspecific” in the sense that the energetic contribution of a given monomer involved in a contact does not depend on its contact partner.

## Results

**Theoretical Analysis.** To use Eqs. 1 and 6 to compute  $Z$ -scores, it is necessary to estimate values for  $\bar{E}$  and  $\sigma_E$ , the average and standard deviation of conformational energies of a given sequence over the ensemble of unfolded conformations. If there are  $M$  unfolded conformations, and unfolded conformation  $j$  corresponds to contact vector,  $\vec{c}_j = (c_{1j}, \dots, c_{Nj})$ , then

$$\bar{E} = \frac{1}{M} \sum_{j=1}^M \left( \sum_{i=1}^N -(h_i c_{ij}) \right) = - \sum_{i=1}^N h_i (\bar{c}_i) = -(\vec{h} \cdot \vec{\bar{c}}) \quad [7]$$

and

$$\sigma_E^2 = \sum_{i=1}^N h_i^2 (\overline{c_i^2} - (\bar{c}_i)^2) = \sum_{i=1}^N h_i^2 (\sigma_c^2)_i, \quad [8]$$

where  $(\bar{c})_i$  and  $(\sigma_c^2)_i$  are the average and variance taken over unfolded conformations of contacts made by monomer  $i$ . Eq. 1 becomes

$$Z = - \frac{\vec{h} \cdot \vec{c} - \vec{h} \cdot \vec{\bar{c}}}{\sqrt{\sum_{i=1}^N h_i^2 (\sigma_c^2)_i}} = - \frac{\vec{h} \cdot (\vec{c} - \vec{\bar{c}})}{\sigma_c \sqrt{\sum_{i=1}^N h_i^2}}, \quad [9]$$

where  $(\sigma_c^2)_i = \sigma_c^2$  was assumed to be independent of  $i$ . Eq. 9 has a clear and important meaning. The sequence,  $\vec{h}$ , not only must be compatible with the target structure,  $\vec{c}$ , but it also must be not compatible with the average unfolded state,  $\vec{\bar{c}}$ . This is best achieved when  $\vec{h}$  is most compatible with  $\vec{c} - \vec{\bar{c}}$ , in the sense that the projection of  $\vec{c} - \vec{\bar{c}}$  on the  $\vec{h}$  direction is maximal. Note that the above expression does not depend on the length of  $\vec{h}$ . Multiplication of all hydrophobicities by a constant is equivalent to a simple change of units of the energy scale and does not affect the  $Z$ -score.

The most negative  $Z$ -score for a given conformation,  $Z_{\text{best}}$ , would correspond to the sequence  $\vec{h}_{\text{best}}$  that minimizes Eq. 9. Direct optimization of this equation is problematic, however, as it involves quantities related to the unfolded state, which is not known *a priori* and can depend itself on the composition of  $\vec{h}$  in a nontrivial way. A simple and useful conformation-dependent upper estimate for  $Z_{\text{best}}$  can nevertheless be obtained if the sequence for a given conformation is taken as

$$\vec{h}_{\text{up}} = \vec{c} - \vec{c}_{\text{up}}, \quad [10]$$

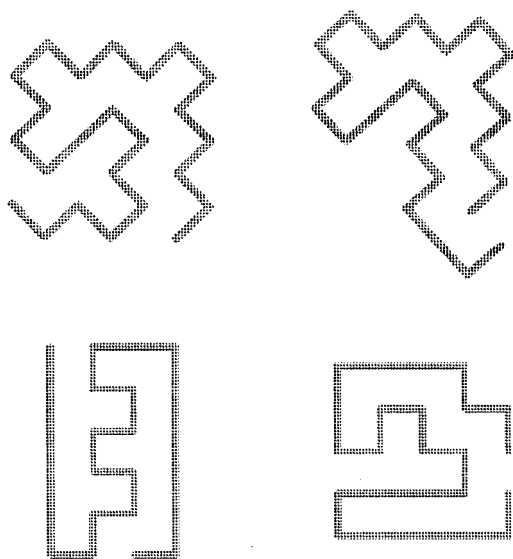
where  $\vec{c}_{\text{up}}$  is the diagonal vector with all components equal to  $C/N$  with  $C = \sum_{i=1}^N c_i$ , and if it is assumed that  $\vec{c}$  is diagonal. This assumption will be appropriate if any possible dependence on sequence position of the average number of contacts made by monomers in the unfolded state (which could arise, for example, from a dependence on individual monomer hydrophobicities) does not deviate  $\vec{c}$  significantly from the direction determined by vector  $\vec{1} = (1, \dots, 1)$ . As  $\vec{h}_{\text{up}}$  is perpendicular to  $\vec{1}$  (i.e.,  $\vec{h}_{\text{up}} \cdot \vec{1} = C - C = 0$ ), the projection of  $\vec{c} - \vec{\bar{c}}$  on  $\vec{h}_{\text{up}} = \vec{c} - \vec{c}_{\text{up}}$  must be equal to this last vector (i.e.,  $(\vec{c} - \vec{\bar{c}}) \cdot (\vec{c} - \vec{c}_{\text{up}}) = (\vec{c} - \vec{c}_{\text{up}})^2$  for any diagonal vector  $\vec{c}$ ) and the corresponding  $Z$ -score, as given by Eq. 9, is independent of the diagonal average unfolded state,  $\vec{\bar{c}}$ .

$$Z_{\text{up}} = - \frac{(\vec{c} - \vec{c}_{\text{up}}) \cdot (\vec{c} - \vec{\bar{c}})}{\sigma_c \sqrt{\sum_{i=1}^N (c_i - C/N)^2}} = - \frac{\sigma}{\sigma_c} \sqrt{N}, \quad [11]$$

where  $\sigma = \sqrt{\sum_{i=1}^N (c_i - C/N)^2}$  is the standard deviation of the number of contacts made by monomers in the native conformation. The scaling with  $\sqrt{N}$  is expected, and it also occurs for  $Z_c$ , Eq. 3, since the entropy scales with  $N$ , i.e.,  $S_0 = N s_0$ , where  $s_0$  is the entropy per monomer.

Note that the upper limit  $Z_{\text{up}}$  will actually be equal to  $Z_{\text{best}}$  if  $\vec{c}$ , for  $\vec{h} = \vec{h}_{\text{up}}$ , happens to be equal to  $\vec{c}_{\text{up}}$  (i.e., if  $\vec{c}$ , the average number of contacts made by each monomer in the unfolded state, is  $C/N$ ). In this hypothetical situation, the sequence taken as in Eq. 10 acquires a very intuitive physical meaning. If a given monomer makes fewer contacts in the target structure than in the unfolded state, then it should be hydrophilic to *destabilize* the unfolded state more than the target conformation. Conversely, if it makes more contacts in the target conformation than in the unfolded state, then it should be hydrophobic to *stabilize* the target conformation more than the unfolded state.

For the more general case, where  $\vec{c}$  is not necessarily equal to  $C/N$ , Eq. 11 shows that for a given conformation with contact standard deviation  $\sigma$ , its best possible  $Z$ -score is at least as negative as  $Z_{\text{up}}$ . For a given  $N$ , structures with abnormally high standard deviations of the number of contacts made by its monomers will therefore correspond to very negative  $Z$ -scores. In other words, conformations with monomers “maximally segregated” between completely buried and completely exposed



**Fig. 1.** Two-dimensional conformations discussed in the present study: **conf1**, Upper Left; **conf2**, Upper Right; **conf3**, Lower Left; and **conf4**, Lower Right. They are all in a regular square lattice and are 24 beads long.

environments are likely to be the most appropriate native structures for the hydrophobic energy function. This inside (more contacts than average)/outside (fewer contacts than average) segregation of monomers in the native conformation is reflected, because of Eq. 10, on the apolar (positive hydrophobicity)/polar (negative hydrophobicity) distribution on the corresponding sequence.

**MC Simulations.** Monte Carlo simulations of a two-dimensional square lattice protein model were performed as a simple test for the presently proposed principle of maximal inside/outside segregation for good native structures. Chains of 24 beads were chosen for this initial study because they can be accommodated inside a region of the lattice with the interesting property of having 12 internal and 12 external sites. Conformation **conf1** (Fig. 1 Upper Left) was generated to fit inside this region. If it were not for chain ends, each bead accommodated on an internal site would make the maximal number of two contacts, while the ones on external sites would make no contacts at all. This would result in an average of 1 contact per monomer with a corresponding standard deviation also of 1. Chain end effects are significant, however, at least for chains of this size. The average and standard deviation of contacts per monomer of conformation **conf1** is  $\bar{c} = 1.08$  and  $\sigma = 0.95$  (Table 1), respectively. An

unexpected rearrangement of **conf1** is **conf2** (Fig. 1 Upper Right), which is less compact (12 contacts instead of 13) and does not fit inside the original region. As it has one of the chain ends making three contacts, however, its contact standard deviation increases to  $\sigma = 1$ , while the average decreases to  $\bar{c} = 1$ . Ideal hydrophobicity vectors for these two conformations (sequences **seq1** and **seq2**) were generated according to Eq. 10, with the value of  $\bar{c}$  taken as unity (Table 1).

A short simulation of **seq2** using the standard Metropolis algorithm (21) with end flips, kinks and crankshaft moves (see, e.g., ref. 22) beginning from a randomly generated initial conformation illustrates how the chain easily folds to the corresponding target conformation in an apparent cooperative two-state transition (Fig. 2 Upper). The simulation temperature is  $T = 0.9$ , which happens to be close to the folding temperature (see Table 1). Energy and number of native contacts,  $Q$ , are plotted every thousand time steps (1 time step equals  $N = 24$  move attempts). The minimal energy is  $E = -24$ , which corresponds to the maximal number of native contacts,  $Q = 12$ . The distribution of the equilibrium conformational ensemble at the same temperature grouped by  $Q$  gives a direct confirmation that the transition is reasonably two-state (Fig. 2 Lower). The distribution was obtained from the average of 10 independent simulations, 10 million time steps each, recorded every 10,000 steps. Results were qualitatively similar for **seq1**, although the transition seemed slightly less cooperative and the contribution of intermediate values of  $Q$  in the equilibrium ensemble at the folding temperature was more pronounced (not shown).

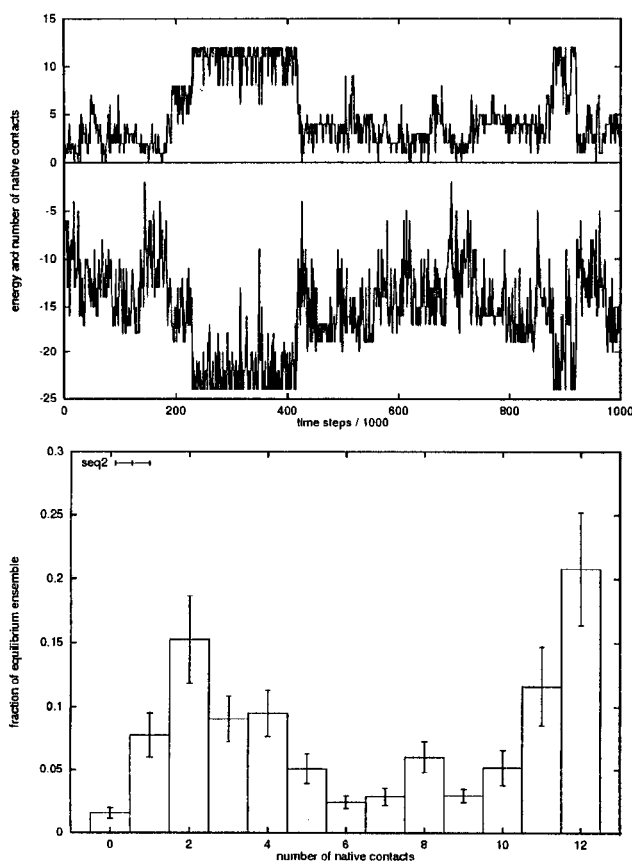
Very long simulations of 100 million time steps, recorded every 10,000 steps, were also performed. At simulation temperature  $T = 1$ , the chain folded and unfolded many times and no other conformation with energy less than or equal to the target conformation was ever found. The histogram technique (14) was used to generate the thermal average of the number of native contacts ( $Q$ ) and the heat capacity ( $C_V = \sigma_E^2/T^2$ ) as a function of temperature (Fig. 3). Representative averages and standard deviations of the same quantities obtained from 10 independent simulations of 10 million time steps are also shown for **seq1** at some temperatures. Good agreement with the curve indicates that the long simulation used by the histogram technique actually sampled an adequate fraction of conformational space.

Heat capacity curves for both **seq1** and **seq2** display the characteristic single sharp peak of cooperative transitions. The maximum of this peak was used to define the folding transition temperature  $T_f$ , which is higher for **seq1** than for **seq2**. The transition of **seq2**, however, as monitored by  $Q$ , seems to be more cooperative than for **seq1**. Note that values of  $T_f$  determined from the  $C_V$  curves agree very well, for **seq1** and **seq2**, with the temperature at the middle of the transition as monitored by the temperature dependence of  $Q$ . This is consistent with a two-state transition and indicates that  $Q$  can be used as an order param-

**Table 1. Conformations and corresponding ideal sequences represented by contact and hydrophobicity vectors, respectively.**

	$\bar{c}$ (upper) and $\bar{h}$ (lower) vectors																$\bar{c}$	$\sigma$	$E$	$T_f$							
<b>conf1</b>	1	0	2	0	2	2	2	2	2	0	2	0	0	2	0	2	0	2	0	2	0	1					
<b>seq1</b>	0	-1	1	-1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	1.08	0.95	-24	0.92	
<b>conf2</b>	1	0	1	0	2	1	2	2	2	0	0	2	0	0	2	0	2	0	2	0	2	3					
<b>seq2</b>	0	-1	0	-1	1	0	1	1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	1	-1	2	1.00	1.00	-24	0.87	
<b>conf3</b>	1	1	1	1	1	0	1	2	2	2	2	2	2	2	1	1	0	1	1	1	1	0	2				
<b>seq3</b>	0	0	0	0	0	-1	0	1	1	1	1	1	1	1	0	0	-1	0	0	0	0	-1	1	1.25	0.66	-18	—
<b>conf4</b>	2	0	1	1	1	0	1	2	2	2	2	2	2	2	1	1	0	1	1	0	2	0	2				
<b>seq4</b>	1	-1	0	0	0	-1	0	1	1	1	1	1	1	1	0	0	-1	0	0	-1	1	-1	1	1.25	0.77	-22	0.71

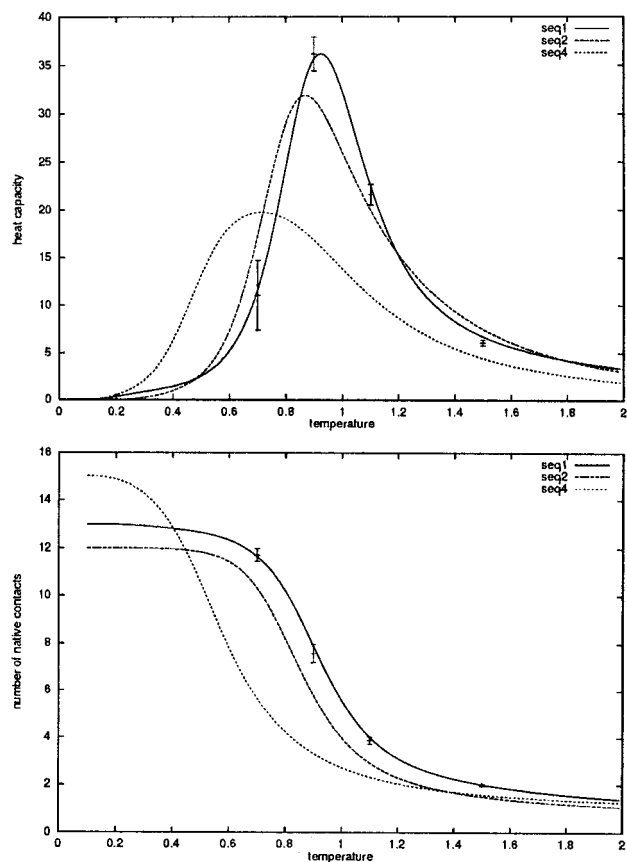
The average number of contacts per monomer,  $\bar{c}$ , contact standard deviation,  $\sigma$ , energy,  $E = -\bar{c}\bar{h}$ , and folding transition temperature determined by the heat capacity peak,  $T_f$ , are also shown. No  $T_f$  was calculated for **seq3** because it does not fold to a single structure (see text).



**Fig. 2.** (Upper) A short Monte Carlo trajectory for **seq2** at  $T = 0.9$ , which is close to its folding temperature. Energy and number of native contacts are displayed for a total time of 1 million time steps, recorded every 1,000 steps. The chain folds cooperatively to its target conformation which corresponds to the energy value of  $-24$  and 12 native contacts. (Lower) The equilibrium ensemble for the same sequence and at the same temperature, grouped according to the number of native contacts. The histogram was constructed from the average over 10 simulations of 10 million time steps run at  $T = 0.9$ . Error bars represent the standard deviation of the mean.

eter. It is interesting to note that the unfolded state has a larger heat capacity than the folded state for both sequences. This result is somehow unexpected because the well known increase in heat capacity upon unfolding of real proteins is traditionally attributed to the ordering of water around exposed apolar residues (23). As there are no explicit water molecules in the present simulations, this result points out the possibility that other factors in addition to solvation can be involved in the experimentally observed heat capacity change.

Conformation **conf3** (Fig. 1 Lower Left) was generated as an example of maximally compact conformation inbedded in a regular region (a  $4 \times 6$  rectangle, in this case). It has 15 contacts but its contact standard deviation is low. Long simulations of the corresponding sequence **seq3** found no conformation with energy lower than the expected  $-18$  but showed that this value corresponded not only to **conf3**, but also to a completely different conformation with the same number of contacts, **conf4** (Fig. 1 Lower Right). As **seq3** has two different conformations with minimal energy, it cannot fold in a protein-like manner. **conf3** and **conf4** have only 5 out of 15 contacts in common, and their contact vectors are not identical. Note that **conf4** has a slightly higher contact standard deviation than **conf3** and, according to the segregation principle, would have a better chance of being the native conformation of a better chosen sequence.



**Fig. 3.** Heat capacity (Upper) and average number of native contacts (Lower) as a function of temperature for **seq1**, **seq2**, and **seq4**. Curves were obtained from a very long simulation of 100 million time steps run at a temperature close to  $T_f$  for each sequence. Independent results obtained from simulations of 10 million time steps run at selected temperatures are shown for **seq1** in the form of the average value over 10 simulations (points) and corresponding standard deviation (error bars).

The sequence obtained from its contact vector, **seq4**, corresponds to a much lower energy,  $-22$ .

The same procedure used for **seq1** and **seq2** suggests that **conf4** is the single global energy minimum for **seq4** but the corresponding temperature dependences of  $C_V$  and  $Q$  are different (Fig. 3). There is a single peak in the  $C_V$  curve, but it is very broad and its maximum occurs at a higher temperature ( $T_f = 0.71$ ) than the mid-point of the transition as monitored by  $Q$ , which is less than 0.6. Taken together, the results for **seq4** show that its folding is much less cooperative and that the transition can hardly be considered as protein-like.

## Discussion

The hydrophobic energy function, upon which this study is based, is no more than a lattice implementation of the physical assumption that conformational free energy of real proteins can be computed from solvent accessible surface areas of different types, such as aliphatic, aromatic and polar, combined with unit free energies of hydration, which can in principle be obtained from partition experiments (24–27). Recent estimates for free energy changes involved in specific pairwise interactions obtained from HPLC experimental data are not inconsistent with this simple scheme (28, 29). A recent decomposition (30) of the Miyazawa and Jernigan potential (31) has also suggested that hydrophobicity can largely explain the statistical distribution of amino acid contacts in protein structures. The conjecture that

the energy can be expressed as in Eq. 6 is particularly interesting because sequence and conformation appear to contribute to the energy on an equal basis. Favorable energies correspond to a large scalar product between two vectors, which can be interpreted as a “compatibility” between sequence and structure.

Sequences and structures are not equivalent, however, as sequences are simple walks in one dimension (each step being a hydrophobicity) while structures are self-avoiding walks in three dimensions. As a consequence, there is an inevitable loss of information when conformations are represented by contact vectors and some contact vectors can represent more than one conformation. It is not clear whether this is evidence that the folding potential felt by real proteins cannot be represented in such a simple form or whether it points out another real physical restriction imposed on protein native structures, as suggested before in the context of maximally compact conformations (20). Some of the intrinsic differences between sequences and conformations are nevertheless reflected in the present vector representation. For example, the set of possible contact vectors is much more restricted than the set of possible hydrophobicity vectors, as most  $N$ -dimensional vectors represent no viable conformation.

The situation is similar to the one described in ref. 20, where it was additionally proposed that the number of sequences encoding a given conformation is proportional to the volume of the Voronoi polytope around its contact vector. This prediction arose from the restriction of conformational space to maximally compact conformations and the explicit assumption that all structures had the same number of surface and core residues (20). This is equivalent to restrict all contact vectors to a single vector length. An analogous proposal for the present model, where polytopes around  $\vec{c}$  are replaced by polytopes around  $\vec{c} - \vec{c}$ , is not valid, however, because the length of  $\vec{c} - \vec{c}$ , even for the simplest case of constant  $\vec{c}$  (i.e., sequences with same composition), depends on  $\sigma$ . A good illustration is provided by **seq3**, which is a vector intended to be parallel to **conf3** -  $\vec{c}$ . As **seq3.conf3** equals **seq3.conf4** it means that the projection of **seq3** on **conf3** -  $\vec{c}$  equals the one on **conf4** -  $\vec{c}$ , which is not parallel to **seq3**. This situation is possible because **conf4** corresponds to a larger  $\sigma$  than **conf3**.

It was implicitly assumed in the present theoretical analysis that monomer hydrophobicities could be tuned arbitrarily to optimize the direction of hydrophobicity vectors. It is clear, however, that the set of possible directions cannot be continuous for real protein sequences, as they are constructed from a limited number of amino acids. In more detailed model, therefore, optimization of Eq. 9 should be restricted to sequences that are compatible with a limited alphabet. Note, however, that such eventual restrictions could probably be at least partly compensated for by small adjustments of the contact vector, which should be continuous in this more realistic situation. In addition, it is important to remember that the alphabet of real amino acids is in itself a product of evolution. The requirement of producing sequences with hydrophobicity vectors sufficiently close to the appropriate directions determined by evolving structures could well have been an important selective pressure.

The results from Monte Carlo simulations demonstrate the possibility of folding protein lattice models with this simple and unspecific hydrophobic energy function when the native conformation is well chosen. The specificity required for the chain to fold to a single conformation does not come from specific interactions between its monomers, but only from the specific pattern of hydrophobicities along its sequence. This result supports, therefore, the basic assumption behind the many enumeration studies performed with the two letter “exact” HP model, which is similar in spirit to the present energy function (reviewed in ref. 19). Complete enumeration of small chains has actually shown that sequences made up of

only two letters, H for hydrophobic and P for polar, are capable of displaying protein-like thermodynamics. In addition, this behavior was found to be dependent on structural details of the native conformation.

Actually, it is interesting to note that the presently proposed principle of inside/outside segregation can help to rationalize some previous results observed for the HP model even though the hydrophobic energy function, from which it was derived, is not identical to the HP function. In addition to the possibility of more than two letters for the hydrophobic function, the two functions do not account for hydrophilic (polar) residues in the same manner. Polar-polar interactions are neutral in the HP model but repulsive according to the hydrophobic function, because of the unfavorable contribution from both residues involved in the contact. In this way, even for the case of an ideal native structure, with all monomers either completely buried or completely exposed, the two-letter alphabet resulting from the hydrophobic function would not be HP. According to the description of  $2 \times 2$  matrices used in ref. 17 the limit hydrophobic function,  $\begin{pmatrix} -1 & 0 \\ 0 & +1 \end{pmatrix}$ , corresponds to an “angle” of 0, while for the HP function,  $\begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}$ , the angle is  $-\arccos(\sqrt{2}/3)$ . A good example of the relation between folding cooperativity and native structure contact variance discussed here is, however, unknowingly provided by figure 24 of ref. 19 for the HP model. The failure of previous attempts to use the stability criterion for HP sequences (19, 32) can also be explained by the small inside/outside segregation of the target maximally compact conformations.

The possibility of a non-maximally compact global energy minimum has been avoided in many studies by making all interactions attractive (e.g., refs. 13 and 33), but this *ad hoc* solution is clearly not consistent with experiments. Real proteins contain hydrophilic residues which interact more favorably with the solvent than with hydrophobic moieties, and therefore their effective interactions with hydrophobic residues cannot be favorable. The present results suggest that this apparently innocent requirement of maximal compactness, in addition to not being rooted on any fundamental physical principle, can be an unnecessary source of artificial complications. Maximally compact conformations are actually poor choices for native structures in the case of square lattice models because most of the monomers make not the maximal or minimum number of contacts but something in between. It has also been noted, even for the case of real proteins, that the assumption of maximal compactness is not well justified. Globular proteins are certainly compact, but their shapes differ significantly from perfect spheres (reviewed in ref. 19). This study suggests that this relative compactness is likely to be an indirect effect of the principle of inside/outside segregation and not of fundamental importance in itself.

Finally, the generality of the theoretical analysis suggests that inside/outside segregation could well be involved in many structural properties observed in real proteins. Although these initial computational results do not permit such generalizations, it is not unreasonable to speculate that segregation can be related to the organization of long chains in independent domains, for example, or even to properties of fibrous proteins, such as the association of long helices to form coiled coils. Further studies involving a systematic exploration of conformational space of different chain sizes, in two and three dimensions, as well as a direct computation of some quantity equivalent to  $\sigma$  in real structures, will be required to elucidate these fundamental questions concerning the generality and applicability of the principle of inside/outside segregation in protein structures.

I am grateful to Fernando Oliveira and Georgios Pappas, Jr. for useful discussions and to Tom Pochapsky and Eugene Shakhnovich for their

critical reading of the manuscript. This study was supported by the Brazilian Conselho Nacional de Pesquisa (CNPq).

1. Goldstein, R., Luthey-Schulten, Z. & Wolynes, P. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 4918–4922.
2. Šali, A., Shakhnovich, E. & Karplus, M. (1994) *Nature (London)* **369**, 248–251.
3. Shakhnovich, E. I. (1994) *Phys. Rev. Lett.* **72**, 3907–3910.
4. Bryngelson, J., Onuchic, J., Succi, N. & Wolynes, P. (1995) *Proteins: Struct. Funct. Genet.* **21**, 167–195.
5. Bowie, T. U., Luthy, R. & Eisenberg, D. (1991) *Science* **253**, 164–169.
6. Gutin, A., Abkevich, V. & Shakhnovich, E. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 3066–3076.
7. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1995) *J. Mol. Biol.* **252**, 460–471.
8. Bryngelson, J. & Wolynes, P. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 7524–7528.
9. Shakhnovich, E. & Gutin, A. (1989) *Biophys. Chem.* **34**, 187–199.
10. Derrida, B. (1981) *Phys. Rev. B* **24**, 2613–2626.
11. Shakhnovich, E. & Gutin, A. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 7195–7199.
12. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994) *J. Chem. Phys.* **101**, 6052–6062.
13. Succi, N. & Onuchic, J. (1994) *J. Chem. Phys.* **101**, 1519–1528.
14. Succi, N. & Onuchic, J. (1995) *J. Chem. Phys.* **103**, 4732–4744.
15. Pereira de Araújo, A. F. & Pochapsky, T. C. (1996) *Folding Design* **1**, 299–314.
16. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. (1997) *Annu. Rev. Phys. Chem.* **48**, 545–600.
17. Pande, V. S., Grosberg, A. & Tanaka, T. (1997) *Biophys. J.* **73**, 3192–3210.
18. Shakhnovich, E. I. (1997) *Curr. Opin. Struct. Biol.* **7**, 29–40.
19. Dill, K., Bronberg, S., Yue, K., Fiebig, K., Yee, D., Thomas, P. D. & Chan, H. S. (1995) *Protein Sci.* **4**, 561–602.
20. Li, H., Tang, C. & Wingreen, N. S. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 4987–4990.
21. Metropolis, N., Rosebluth, A., Rosebluth, M. & Teller, A. (1953) *J. Chem. Phys.* **21**, 1087–1092.
22. Betancourt, M. & Onuchic, J. (1995) *J. Chem. Phys.* **103**, 773–787.
23. Privalov, P. L. (1992) in *Protein Folding*, ed. Creighton, T. E. (Freeman, New York).
24. Murphy, K. P., Bhakuni, V., Xie, D. & Freire, E. (1992) *J. Mol. Biol.* **227**, 293–306.
25. Makhatazde, G. I. & Privalov, P. L. (1993) *J. Mol. Biol.* **232**, 639–659.
26. Makhatazde, G. I. & Privalov, P. L. (1993) *J. Mol. Biol.* **232**, 660–679.
27. Chang, H. S. & Dill, K. A. (1997) *Annu. Rev. Biophys. Biomol. Struct.* **26**, 425–459.
28. Pochapsky, T. C. & Gopen, Q. (1992) *Protein Sci.* **1**, 786–795.
29. Pereira de Araújo, A. F., Pochapsky, T. C. & Joughin, B. (1999) *Biophys. J.* **76**, 2319–2328.
30. Li, H., Tang, C. & Wingreen, N. S. (1997) *Phys. Rev. Lett.* **79**, 765–768.
31. Miyazawa, S. & Jerningan, R. (1985) *Macromolecules* **18**, 534–552.
32. Yue, K., Fiebig, K. M., Thomas, P. D., Chang, H. S., Shakhnovich, E. I. & Dill, K. A. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 325–329.
33. Li, H., Helling, R., Tang, C. & Wingreen, N. S. (1996) *Science* **273**, 666–669.