Relevance of structural segregation and chain compaction for the thermodynamics of folding of a hydrophobic protein model

Marco Aurélio A. Barbosa
Instituto de Física, Universidade de Brasília, Brasília-DF 70910-900, Brazil

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 3 February 2003; published 21 May 2003)

The relevance of inside-outside segregation and chain compaction for the thermodynamics of folding of a hydrophobic protein model is probed by complete enumeration of two-dimensional chains of up to 18 monomers in the square lattice. The exact computation of *Z* scores for uniquely designed sequences confirms that *Z* tends to decrease linearly with $\sigma\sqrt{N}$, as previously suggested by theoretical analysis and Monte Carlo simulations, where σ, the standard deviation of the number of contacts made by different monomers in the target structure, is a measure of structural segregation and *N* is the chain length. The probability that the target conformation is indeed the unique global energy minimum of the designed sequence is found to increase dramatically with σ, approaching unity at maximal segregation. However, due to the huge number of conformations with sub-maximal values of σ, which correspond to intermediate, only mildly discriminative, values of *Z*, in addition to significant oscillations of *Z* around its estimated value, the probability that a correctly designed sequence corresponds to a maximally segregated conformation is small. This behavior of *Z* also explains the observed relation between σ and different measures of folding cooperativity of correctly designed sequences.

DOI: 10.1103/PhysRevE.67.051919 PACS number(s): 87.15.Cc, 87.10.+e

I. INTRODUCTION

A recent theoretical analysis, corroborated by Monte Carlo simulations of lattice protein models, has suggested that structural segregation as measured by σ, the standard deviation of the number of contacts made by each monomer in a given structure, is an important property of appropriate native conformations for a simple energy function intended to mimic the hydrophobic effect [1–4]. This hydrophobic energy function, which has also been used by other groups in different studies [5–8], adopts the convenient form of a scalar product between two vectors in *N*-dimensional space, with *N* being the number of monomers:

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

where $\vec{h}=\{h_1,\ldots,h_N\}$ represents the sequence of hydrophobicities along the chain and $\vec{c}=\{c_1,\ldots,c_N\}$ represents the number of contacts made by each monomer in a given conformation. Hydrophobic monomers (positive hydrophobicity) tend in this way to make contacts, hiding from the solvent, while the reverse is true for hydrophilic monomers (negative hydrophobicity). Note that the sum over monomers of this equation can be replaced by the more familiar sum over contacts if the energetic contribution of each contact, e.g., between monomers *i* and *j*, is taken simply as $-(h_i+h_j)$.

The combination of the hydrophobic energy function with the *Z*-score criterion for proteinlike folding behavior implies that the absolute value of the following quantity should be large:

$$Z(\vec{h},\vec{c^*})=\frac{E^*-\overline{E}}{\sigma_E}=\frac{-\vec{h}\cdot(\vec{c^*}-\vec{\overline{c}})}{\sigma_c\sqrt{\sum_{i=1}^{N} h_i^2}}, \tag{2}$$

where $E^*=-\vec{h}\cdot\vec{c^*}$ is the putative native energy, i.e., the energy of sequence $\vec{h}$ in its putative native structure $\vec{c^*}$. The average energy of sequence $\vec{h}$ over unfolded conformations, $\overline{E}$, and the corresponding standard deviation σ_E were estimated in the expression above as

$$\overline{E}=-\sum_{i=1}^{N} h_i(\overline{c})_i=-\vec{h}\cdot\vec{\overline{c}} \tag{3}$$

and

$$\sigma_E=\sqrt{\sum_{i=1}^{N} h_i^2(\sigma_c^2)_i}, \tag{4}$$

where $(\overline{c})_i$ is the average over unfolded conformations of the number of contacts made by monomer *i* and $(\sigma_c^2)_i=\sigma_c^2$ is the corresponding variance, which was further assumed to be independent of *i*. The difference vector $\vec{c^*}-\vec{\overline{c}}$ must in this way have a large projection upon the direction defined by the unitary vector $(1/|\vec{h}|)\vec{h}$, where $|\vec{h}|=\Sigma_{i=1}^{N}h_i^2$ is the length of $\vec{h}$ [1].

*Electronic address: aaraujo@unb.br

If the average unfolded state vector is assumed to be diagonal, i.e., if $\vec{\bar{c}}=\{\bar{c},\ldots,\bar{c}\}$ with $\bar{c}$ being the average number of contacts made by each monomer in the unfolded state, then a useful conformation-dependent upper limit for the best possible Z score for a putative native contact vector $\vec{c^*}$ can be obtained by choosing $\vec{h}$, the corresponding sequence, simply as

$$\vec{h}=\vec{c^*}-\vec{c^*_D}, \tag{5}$$

where $\vec{c^*_D}=\{C/N,\ldots,C/N\}$ is the projection of $\vec{c^*}$ upon the diagonal direction defined by vector $\vec{1}=\{1,\ldots,1\}$ and $C=\Sigma_{i=1}^{N}c_i^*$ is twice the total number of contacts in the putative native structure. With this choice of sequence the putative native energy is given by

$$E^*=-(\vec{c^*}\cdot\vec{c^*}-\vec{c^*_D}\cdot\vec{c^*})=-\sum_i (c_i^*)^2-\sum_i \frac{C}{N}c_i^*=-N\sigma^2, \tag{6}$$

and the estimate for standard deviation over unfolded conformations [Eq. (4)] becomes

$$\sigma_E=\sigma_c\sqrt{\sum_{i=1}^{N}\left(c_i^*-\frac{C}{N}\right)^2}=\sigma_c\sigma\sqrt{N}, \tag{7}$$

where $\sigma=\sqrt{(1/N)\Sigma_{i=1}^{N}(c_i^*-C/N)^2}$ is the standard deviation of the number of contacts made by monomers *in the native conformation*. Since the sequence is perpendicular to the diagonal direction the average energy over unfolded conformations is zero for any diagonal vector $\vec{\bar{c}}$ and Z is given simply by [1]

$$Z=-\frac{\sigma}{\sigma_c}\sqrt{N}. \tag{8}$$

The dependence of Z on σ implies that structurally segregated conformations, having most of their monomers evenly distributed between completely buried positions, making the maximal number of contacts, and completely exposed positions, making no contacts at all, would have more negative Z scores and, therefore, would have a better chance to reproduce proteinlike folding behavior. This expectation has been corroborated by the successful design of sequences intended to fold to structurally segregated conformations in square [1,2,4] and cubic [3] lattices. In addition, this dependence of Z on segregation was suggested to be partially responsible for previous unsuccessful attempts to design sequences to maximally compact conformations with simple energy functions intended to mimic the hydrophobic effect [9,10], unless pair-specific segregation terms [7,11] or nearest-neighbor constraints [12] were included, since maximally compact conformations in square and cubic lattices happen to have low values of σ. It is also interesting to note that an independent mean-field analysis has recently concluded that contact vector degeneracy, i.e., the number of different structures corresponding to the same vector, should decrease with σ [13], further corroborating the hypothesis that structurally segregated conformations should have a better chance of reproducing proteinlike folding behavior for energy functions dependent on contact vector information alone.

TABLE I. For each chain length, expressed by the number of monomers N, column A shows the total number of structures, column B shows the number of structures that are uniquely determined by their contact vectors, and column C shows the number of structures for which sequences are successfully designed according to Eq. (5), i.e., the target conformation actually corresponds to the nondegenerate global energy minimum of the designed sequence. Column D was taken directly from Table II of Ref. [17] and shows the number of conformations uniquely determined by their contact matrices, while column E displays the ratio between columns B and D, i.e., the fraction of conformations uniquely determined by their contact matrices, which are also uniquely determined by their contact vectors.

N	A	B	C	D	E
3	2	0	0		
4	5	1	1		
5	13	0	0		
6	36	4	3		
7	98	2	2		
8	272	23	11		
9	740	25	17		
10	2034	100	48		
11	5513	154	96	154	1.000
12	15037	509	191	519	0.981
13	40617	868	418	898	0.967
14	110188	2587	881	2836	0.912
15	296806	4494	1786	4954	0.907
16	802075	13018	3722	15048	0.865
17	2155667	23347	8837	26494	0.881
18	5808335	65340	17499	77635	0.842

In the present study we use complete enumeration of chains of up to 18 monomers in the square lattice to compute the exact dependence of Z on segregation, as measured by σ, and compaction, as measured by C/N, for sequences designed by Eq. (5) to all structures that are uniquely determined by their contact vectors. We find that our estimates for the unfolded state are quite accurate and that the overall dependence of Z on σ and N is reasonably well predicted by Eq. (8). There are significant fluctuations around the estimate, however, particularly at submaximal values of σ, which in addition tend to correspond to intermediate, weakly predictive values of Z. This behavior can be used to explain the dependence of the probability of successful design on structural segregation and the relation between this segregation and different measures of cooperativity for successfully designed sequences.

II. RESULTS AND DISCUSSION

All present enumeration results were obtained by a recursive algorithm that counts all self-avoiding walks of $N-1$ steps in the square lattice not related by rotation or reflection. The obtained total number of walks, or conformations, for N up to 18, is shown in column A of Table I and agrees exactly

with previously published data [14–16]. Not all of these conformations can be used as native structures, however. Since the conformation enters the hydrophobic energy function exclusively through its contact vector, it is clear that degenerate contact vectors, i.e., vectors that represent more than one conformation, cannot correspond to the global, nondegenerate, energy minimum of any sequence. The number of structures that conform to this obvious criterion, i.e., the number of ''potentially encodable'' structures [17], is shown on column *B* of Table I for different values of *N*. Column *C* of the same table shows how many of these potentially encodable structures actually correspond to the nondegenerate global energy minimum for the sequence designed according to Eq. (5), i.e., the number of potentially encodable structures for which sequence design was successful. Single exponential fits for all three sets of conformations as a function of *N* are shown in Fig. 1.

It is clear that the number of potentially encodable conformations for the hydrophobic energy function is much smaller than the total number of conformations (65 340 against 5 808 335 for 18mers, or about two orders of magnitude) and that this difference increases with chain length. In other words, it is most likely that among contact vectors generated by self-avoiding walks, an immense number of them is degenerate, in the sense that they represent more than one possible conformation. This fact could suggest that contact vectors would not be able to encode the amount of information required for folding simulations since a huge number of conformations cannot be uniquely defined by them. It is important to notice, however, that an appropriate folding model is not required to uniquely determine, in terms of energy, all possible conformations, but only native structures. Most conformations that are not uniquely determined by their contact vectors have a small number of contacts and would not be appropriate models for native structures of globular proteins in any case.

Numbers shown in column *B* of Table I are actually always smaller or equal to the number of potentially encodable conformations for the energy functions that can distinguish between contact matrices instead of contact vectors, as seen in column *D* of the same table, which was transcribed from Table II of Ref. [17]. The explanation for this observation is quite simple. If two conformations have different contact vectors they must necessarily have different contact matrices, since the components of the contact vector are uniquely determined by the sums over columns or rows of the contact matrix. The reverse is not true, however, since different matrices can correspond to the same vector [15,13]. This point is particularly important from the perspective of information theory, since it is closely related to the small amount of information contained in contact vectors when compared to contact matrices. The amount of information contained in contact vectors cannot be larger than $N\log_2(z+1)$ bits, where z is the maximal number of contacts a monomer can make, and is much smaller than the $N(N-1)/2$ bits of information required to define a contact matrix. In particular, the maximal amount of information that can be contained in contact vectors increases only linearly with chain length and, therefore, is compatible with the amount of information contained in

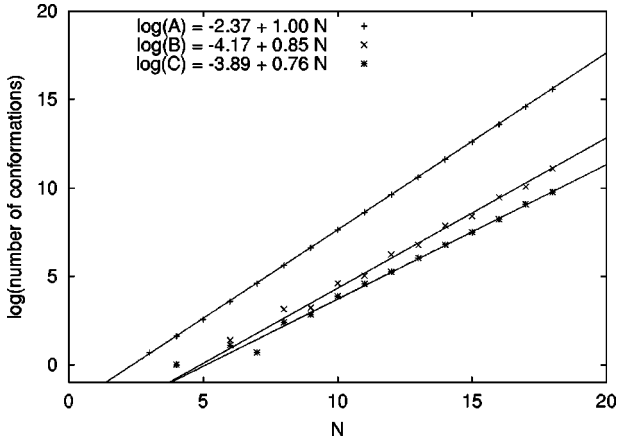

FIG. 1. Exponential dependence on chain length, as measured by the number of monomers, *N*, of the number of all possible conformations (A), the number conformations uniquely determined by their contact vectors (B), and the number of successfully designed conformations (C). The points represent the numbers taken directly from the corresponding columns of Table I and the lines are single exponential fits to the data. All quantities are adimensional.

sequences of *N* letters chosen from a fixed alphabet of *L* letter types, with $L \ll N$, as is the case for real amino acid sequences. It is quite interesting, therefore, that a significant fraction of the conformations uniquely determined by their contact matrices is also uniquely determined by their contact vectors (e.g., $154/154=1$ for $N=11$, $65\,341/77\,635=0.84$ for $N=18$), as seen in column *E* of Table I. Even considering that this fraction tends to decrease with chain length, it becomes apparent that much of the contact matrix extra information might be redundant and not useful in the specific task of recognizing unique native structures among all alternative conformations.

Table II shows the distribution of conformations of 18mers grouped according to segregation and compaction. Different rows correspond to different compactions, labeled by C/N, as well as C, where $C=\Sigma_i c_i$ is twice the total number of contacts, while different columns correspond to different segregations, with σ intervals 0.1 units wide labeled by their inferior limits. Each position of the table shows up to three numbers. The first number is the total number of conformations with the corresponding degrees of segregation and compaction. If this first number is different from zero, then a second number indicates how many of them are potentially encodable, i.e., uniquely determined by their contact vectors. If this second number is also different from zero then a third number indicates how many conformations are successfully designed, in the sense that they actually correspond to the nondegenerate global energy minimum of the sequence designed according to Eq. (5). Note that 646 909 conformations have no contacts at all while 3 208 721 (55%) have two contacts or less (i.e., $C \leqslant 4$ and $C/N \leqslant 0.222$), and only three of them are uniquely determined by their contact vectors. For a given C/N, the distribution tends to be sharply peaked around intermediate values of σ, being consistent with a product between the rapidly increasing

TABLE II. Conformational space of 18mers obtained by complete enumeration. All quantities are adimensional. For each segregation, labeled by the inferior limits of σ intervals 0.1 units wide on the first row, and compaction, labeled by C and C/N on the first column, the table shows up to three numbers. The first number is the total number of conformations with the corresponding degrees of segregation and compaction. If this first number is different from zero, then a second number indicates how many of the total number of conformations are uniquely determined by their contact vectors. If this second number is also different from zero, then a third number indicates how many conformations are successfully designed according to Eq. (5).

$\sigma\rightarrow$ $C\downarrow$ (C/N)	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	Total
0 (0.000)	646909 0	0	0	0	0	0	0	0	0	0	0	0	646909 0
2 (0.111)	0	0	0	1221134 0	0	0	0	0	0	0	0	0	1221134 0
4 (0.222)	0	0	0	0	1162320 3 0	178358 0	0	0	0	0	0	0	1340678 3 0
6 (0.333)	0	0	0	0	633475 98 0	390502 82 0	46232 2 0	4678 0	0	0	0	0	1074887 182 0
8 (0.444)	0	0	0	0	208168 322 0	369988 872 0	124670 312 8	23280 52 0	3716 6 2	0	0	0	729822 1564 10
10 (0.556)	0	0	0	0	37243 220 0	160714 1544 42	159285 1764 419	48324 692 238	15239 147 86	140 0	0	0	420945 4367 785
12 (0.667)	0	0	0	0	3211 133 0	33298 1912 0	88970 5729 313	64780 5226 1024	31034 2955 1187	993 100 92	2 2 2	0	222288 16057 2618
14 (0.778)	0	0	0	0	108 5 0	2140 212 0	16673 1883 70	61474 7566 2213	10752 1902 1148	3664 708 640	14 8 8	0	94825 12284 4079
16 (0.889)	0	0	0	2 2 0	50 50 0	797 537 0	6138 2430 76	13310 5305 840	17077 8601 3256	2434 1612 1298	8 8 8	0	39816 18545 5478
18 (1.000)	0	0	0	0	0	78 74 0	1573 917 8	4360 3116 564	7461 5386 2255	1398 1142 900	484 456 450	4 4 4	15358 11095 4181
20 (1.111)	0	0	0	0	0	32 32 0	236 220 0	459 365 78	820 532 188	126 94 82	0	0	1673 1243 348
Total	646909 0	0	0	1221136 2 0	2044575 831 0	1135907 5265 42	443777 13257 894	220665 22322 4957	86099 19529 8122	8755 3656 3012	508 474 468	4 4 4	5808335 65340 17499

$[(N-2)$-dimensional] associated volume inside the space of contact vectors with a rapidly decreasing conformational density. Qualitatively similar results were obtained for smaller chains.

The two structural parameters under investigation, structural segregation, and compaction, can be associated with perpendicular directions in the space of contact vectors, or in its two-dimensional sequence-structure diagram (SS-diagram) representation [2], since the projection of $\vec{c}$ upon the diagonal is proportional to C/N, while the distance between $\vec{c}$ and the diagonal is proportional to σ. The predicted relation between σ and Z can actually be understood from the fact that segregated contact vectors point out of the densely populated diagonal region and have in this way a better chance of being clearly distinguished from vectors representing the vast majority of alternative conformations. Table II clearly shows, however, that structural segregation and compaction are to some extent correlated for small chains and that maximally segregated conformations are also significantly, although not maximally, compact. For example,

95% (484 out of 508) of the conformations with segregation $\sigma \geqslant 1$ also have compaction $C/N \geqslant 1$. A different behavior should be expected, however, for longer chains since in the limit of infinite length maximally compact conformations would have $C/N=2$ and $\sigma=0$, since all monomers would make two contacts, while maximally segregated conformations would have $C/N=1$ and $\sigma=1$. Note that the fact that long protein molecules do not fold into single large globules but are divided into structural domains might reflect a preponderance of segregation upon compaction on the selection of appropriate native structures.

Z scores were computed exactly for sequences designed to all conformations uniquely determined by their contact vectors (column B of Table I) with unfolded state average energy and standard deviation for each sequence computed over all conformations (column A of Table I) taken as equally likely, i.e., corresponding to the unfolded state at infinite temperature. The energy of the native structure was included in the averaging since its contribution is in this case negligible. Alternatively, unfolded state averages could have been computed at a finite temperature, with different conformations weighted by their Boltzmann factors. This second approach, however, requires some arbitrary choices, such as the temperature, the definition of the macroscopic folded state, i.e., the native structure, and possibly other sufficiently similar conformations which would have to be excluded from the averaging, and also a normalization scheme to put the temperature for all sequences on the same energy scale. Since the Z score at infinite temperature already provides a characterization of the density of states as a whole without these unnecessary complications, we consider the first approach more appropriate for the purposes of the present study.

Figure 2(a) shows the unfolded average energy and standard deviation for sequences designed according to Eq. (5) to all 65 340 18mer conformations uniquely defined by their contact vectors, as well as their global energy minima. A single point might correspond to more than one conformation. The expression for the energy of the target conformation, given by Eq. (6), is also plotted in the figure. Points on this line correspond to sequences for which the target conformation has the minimal energy, although the possibility of others conformations with the same energy is not excluded. Points below the line, abundant at small values of σ, show that the global energy minimum is lower than the energy of the target conformation and sequence design was not successful. The figure also shows that the estimates for the average and standard deviation over unfolded conformations are reasonable since directly computed average energies are always close to zero while the standard deviation appears to increase linearly with σ. The value of σ_c obtained from Eq. (7) with a linear fit to the data, 0.5037 ± 0.0001, agrees extremely well with the corresponding exact quantity computed directly from the enumeration, 0.506 912, although in this last case there is some dependence on monomer position along the sequence, with pronounced minima at the positions next to the extremities (Fig. 3). These minima can probably be attributed to trivial end effects since these monomers have half the number of nearest neighbors when compared to monomers in the middle of the chain. For the extremities this effect would be compensated by the possible formation of up to three contacts, instead of only two.

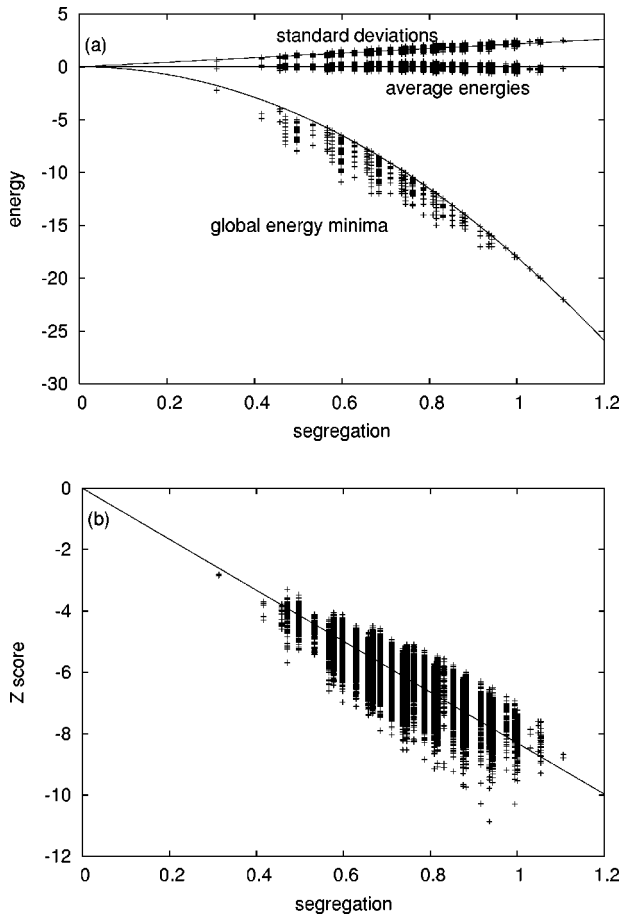

FIG. 2. (a) Dependence on segregation, as measured by σ values, of the average energy and standard deviation over the unfolded state ensemble as well as the actual global energy minimum of sequences designed to all conformations uniquely determined by their contact vectors. The expected zero value for the average unfolded state energy and the linear dependence for the standard deviation are also shown. The parabolic line shows the exact dependence of the energy of the target conformation on σ, $E^* = -N\sigma^2$, with $N=18$. Global minima below this line indicate unsuccessful design. (b) The Z score as a function of segregation, as measured by σ, computed directly from the data shown in (a) and the theoretically predicted linear dependence, $Z=(\sqrt{N}/\sigma_c)\sigma$, with $N=18$ and $\sigma_c=0.51$. Z score and segregation are adimensional. The energy scale is defined by the numerical values of the hydrophobicities along the sequences.

The dependence of Z on σ, computed from the data shown in Fig. 2(a), is shown in Fig. 2(b). Although Z clearly tends to become more negative as σ increases, with a Pearson's correlation coefficient of -0.82, there are strong oscillations around the predicted linear dependence, particularly for intermediate segregations. As a consequence, the most negative values of Z do not necessarily correspond to maximally segregated conformations, as would be expected from the expression given by Eq. (8), which is also plotted in the figure. These significant oscillations on Z arise as a conse-

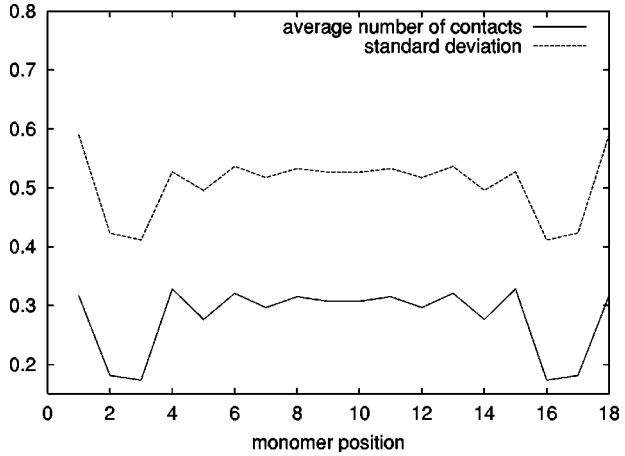

FIG. 3. Average number of contacts made by each monomer in the ensemble of all conformations of 18 monomers, $(\overline{c})_i$, and corresponding standard deviation $(\sigma_c)_i$ as a function of monomer position along the sequence *i*.

quence of oscillations on σ_E, the denominator of Eq. (2), which ultimately result from the approximations made during the derivation of Eq. (4) like, for example, the neglect of the dependence of σ_c on monomer position, or, more importantly, the implicit assumption that the numbers of contacts made by different monomers along the chain are statistically independent. Qualitatively similar results were obtained for smaller values of *N*.

The predicted dependence of *Z* on chain length is verified in Fig. 4, where two related plots summarize the information obtained for $N=12$ to $N=18$. Figure 4(a) shows that the slope of the linear dependence of *Z* on σ, shown in Fig. 2(b) for $N=18$, depends itself on chain length as predicted by Eq. (8). Points representing the slopes of straight lines, $Z=-A\sigma$, were obtained by linear fits to the data and agree very well with the predicted value, $A=(1/\sigma_c)\sqrt{N}$, with $\sigma_c=0.5$, shown as a curve in the same figure. Exact computation of σ_c for different chain lengths shows that σ_c actually tends to increase slightly with *N*, from 0.485 915, for $N=12$, to 0.506 912, for $N=18$. Figure 4(b) shows the quantity $(Z^2\sigma_c^2)/\sigma^2$, which should be equal to *N* according to Eq. (8), plotted as a function of the actual *N*. No adjustment is involved. Each point and error bar corresponds, respectively, to the average and standard deviation of this quantity over all potentially encodable conformations of a given chain length. It is clear that the points tend to lie on the diagonal, as expected, but the size of the error bars indicate, again, the presence of strong fluctuations around the predicted value. Note that the linear dependence of *Z* on $\sqrt{N}$ does not imply, by itself, any advantage for longer chains regarding sequence design since the same dependence is expected for the critical value Z_c, below which proteinlike folding behavior is likely to occur [1].

Having corroborated the overall dependence of *Z* predicted by Eq. (8) both on σ and *N*, we then investigated the relation between structural segregation and successful sequence design using conformations with $N=18$. Points of Fig. 2(b) corresponding to successful and unsuccessful design are shown in Figs. 5(a) and 5(b), respectively. For the largest values of σ ($\sigma>1$) design is always successful, in the sense that the target conformation is actually the global energy minimum of its sequence, designed according to Eq. 5, while for low values of σ ($\sigma<0.6$) this is never the case. At intermediate values of σ, however, which also correspond to intermediate values of *Z*, sequence design can be either successful or unsuccessful. The dramatic increase in the probability of successful design with σ, approaching unity for maximally segregated conformations ($\sigma\geqslant 1$), is shown by curve *A* in Fig. 6(a). Note, however, that a huge number of appropriate conformations corresponding to submaximal values of σ cannot be selected out of their inappropriate counterparts by a segregation criterion. For example, although the probability of successful design for an arbitrary conformation of $\sigma\approx 0.8$ is only around 0.4, this is the most

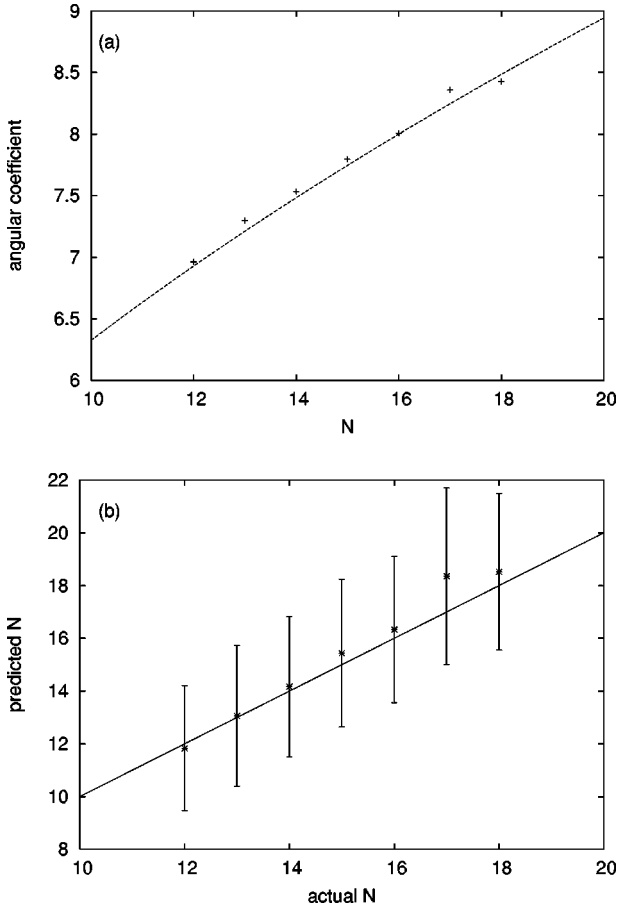

FIG. 4. (a) Dependence on chain length of the slope of the linear dependence of *Z* on σ. Each point represents the angular coefficient *A* of a straight line, $Z=-A\sigma$, fitted to all potentially encodable conformations for each value of *N*. The curve shows the dependence of *A* on *N* predicted by Eq. (8), $A=(1/\sigma_c)\sqrt{N}$, with $\sigma_c=0.5$. (b) Relationship between *N* predicted by Eq. (8), $N=(\sigma_c^2 Z^2)/\sigma^2$, and the actual value of *N*, for the same sets of conformations. Each point and error bar corresponds, respectively, to the average and standard deviation of this quantity over all potentially encodable conformations of a given chain length. σ_c was computed exactly for each chain length. All quantities are adimensional.

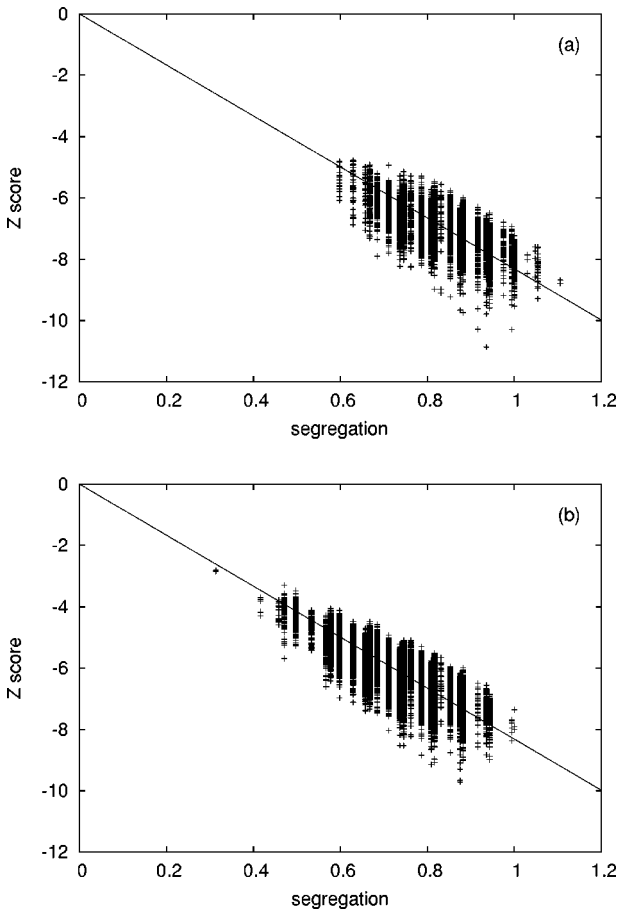

FIG. 5. Same data shown in Fig. 2(b) grouped according to successful design (a) and unsuccessful design (b).

likely segregation value of an arbitrary successfully designed conformation, as seen in curve *B* of the same figure. In other words, σ maximization is a powerful criterion for selecting appropriate native conformations for the hydrophobic energy function, as has been done previously both in two and three dimensions [1–4], but if conformations were to be selected by a different hypothetical mechanism able to sample appropriate conformations uniformly, as it is likely to be the case of natural selection acting on real proteins, then they would tend to have large but not maximal values of σ.

It is already apparent from Fig. 5 that even a perfect correlation between σ and Z would not result in a great improvement in the discrimination of successfully designed conformations at submaximal values of σ because the corresponding intermediate values of Z are themselves not discriminative. Figure 6(b) shows that the relation between Z score and successful sequence design is actually very similar to the one observed for σ. Very good Z score values, e.g., $Z<-8$, correspond to a high probability of successful design (curve *A*) but the vast majority of successfully designed sequences correspond to intermediate values of Z, $-6>Z>-8$ (curve *B*), for which the probability of successful design is around 0.5. It is interesting to note, incidentally, that σ actually appears to be a slightly better indicator of successful design than Z itself in this particular ensemble of 18mers in the square lattice. All conformations with $\sigma>1$, for example, happen to be successfully designed but this could not be inferred from their values of Z [see Figs. 5(a) and 5(b)]. Figure 6(c) shows that if conformations are grouped according to compaction instead of segregation, the probability of successful design is never higher than around 0.4 and that it actually displays a slight decrease at maximal compactness (curve *A*). Successfully designed conformations, however, tend to be significantly compact, with a broad peak of probability around $C/N=0.889$, corresponding to $C=18$ or 9 contacts (curve *B*).

We have also investigated the possibility that structural segregation, although not capable of completely separating successfully designed sequences from their unsuccessfully designed counterparts, could be useful in the identification of the majority good folders among successfully designed sequences. Two folding quality parameters computable from the heat capacity curves of all successfully designed sequences were investigated. First, the cooperativity parameter of Thirumalai and co-workers [18], Ω_c, with the energy itself [or, more precisely, $1-E(T)/E^*$] taken as the reaction coordinate, i.e., the quantity

$$\Omega_c=\frac{T_f^2 C_V(T_f)}{(-E^*)\Delta T}, \tag{9}$$

where T_f is the temperature of maximal heat capacity, $C_V(T)=dE(T)/dT$, and ΔT is the width of the heat capacity peak at its half-maximal height. The factor $(-1/E^*)$ results from normalization of the reaction coordinate, in a way that it will vary from $1-E^*/E^*=0$ at $T=0$ (folded state) to $1-0/E^*=1$ at $T=\infty$ (unfolded state). The heat capacity curve for each sequence was calculated exactly from the enumeration of all possible conformations.

Figure 7(a) shows how Ω_c depends on σ for all successfully designed 18mer sequences. It is again apparent that there is a tendency of Ω_c to increase with σ, since for $\sigma\approx0.6$, corresponding to the least segregated structures for which sequence design is successful, Ω_c varies from almost 0 to around 1, while for $\sigma\approx1.1$, corresponding to the maximally segregated conformations, there are no sequences with $\Omega_c<2$ and some sequences have this parameter above 3, among the highest among all sequences. For intermediate values of segregation, however, there are strong oscillations in cooperativity. For σ between 0.8 and 0.9, for example, Ω_c can have any value between 0 and 3. The Pearson's correlation coefficient between Ω_c and σ for 17 499 successfully designed 18mer conformations is 0.182. The probability that this value results from pure chance is essentially zero, since the number of points is large, but it is clear that the correlation is not strong. The main reason for this rather weak correlation between Ω_c and σ appears to be not the correlation between σ and Z (-0.69 for the same set of points), but actually the mild correlation between Ω_c and Z itself (-0.398), as seen in Fig. 7(b).

The cooperativity parameter Ω_c indicates how abruptly the reaction coordinate changes from the native state value to the unfolded state value as a function of temperature. A significantly more stringent criterion for folding quality has

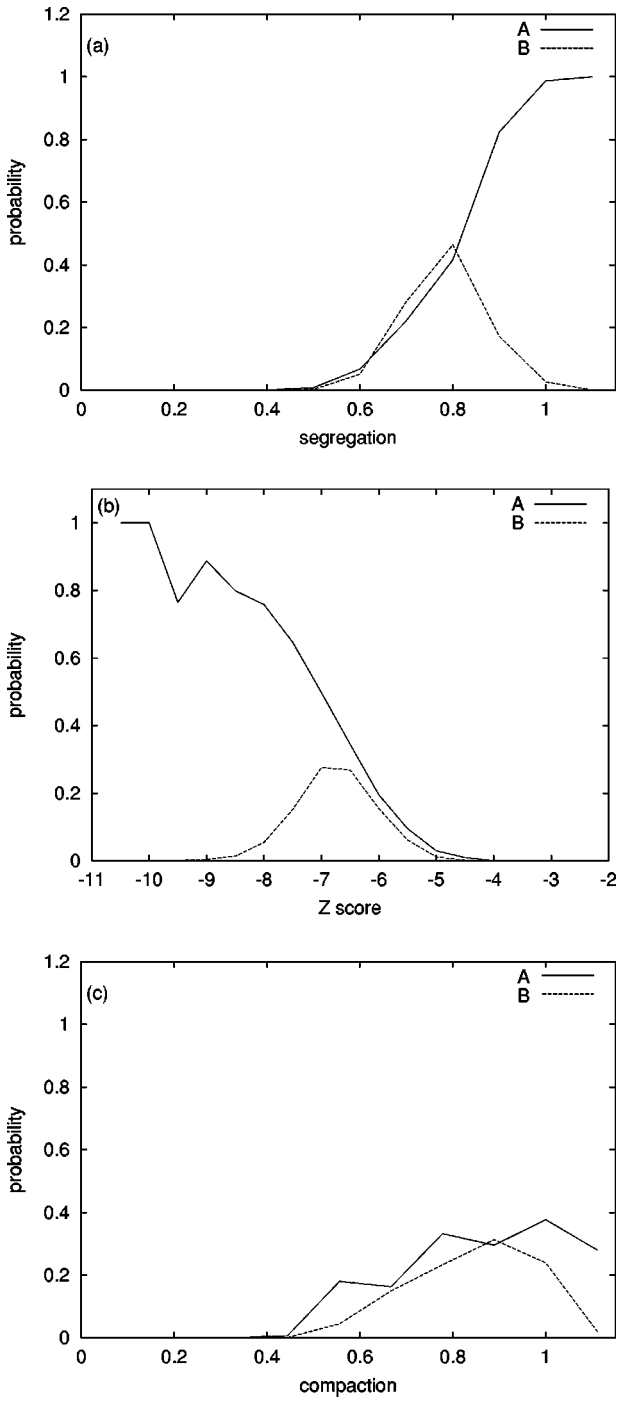

FIG. 6. (a) Probability of successful design as a function of structural segregation, as measured by σ (curve A), and probability that an arbitrary successfully designed conformation will have a given value of σ (curve B). (b) Probability of successful design as a function of Z score (curve A) and probability that an arbitrary successfully designed conformation will have a given value of Z (curve B). (c) Probability of successful design as a function of compaction, as measured by C/N (curve A), and probability that an arbitrary successfully designed conformation will have a given value of C/N (curve B). All quantities are adimensional.

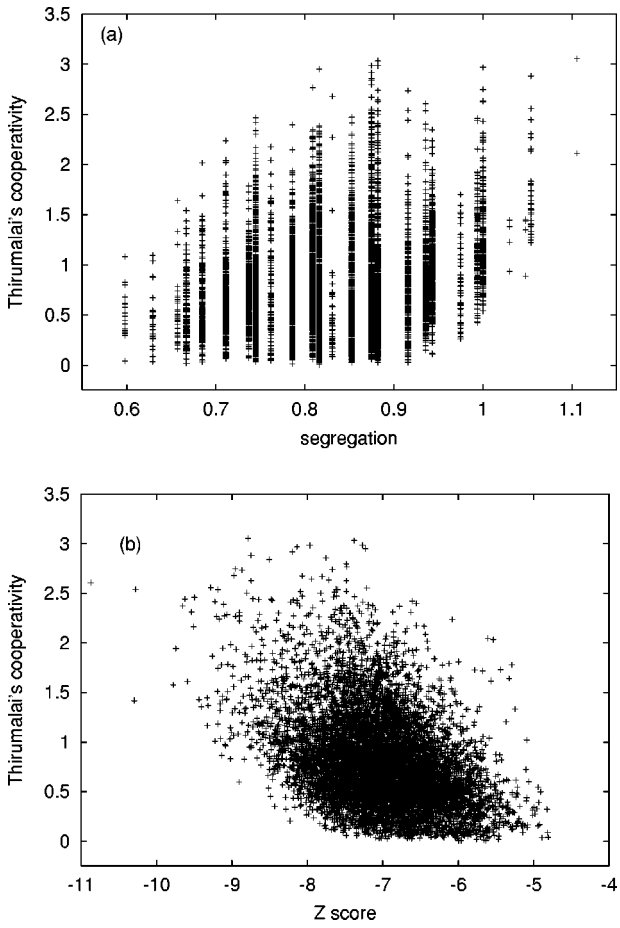

FIG. 7. Relation between folding cooperativity, as measured by Thirumalai's Ω_c parameter, and structural segregation, as measured by σ (correlation coefficient of 0.182) (a) and relation between Ω_c and Z (correlation coefficient of -0.398) (b). All quantities are adimensional. The correlation coefficient between σ and Z for the same set of points is -0.69.

been recently proposed by Chan and co-workers [16,19,20], according to which the energy and associated heat capacity curves of model proteins are compared directly to the enthalpy and associated heat capacities of real proteins, for which the (calorimetrically measured) heat absorbed during thermal unfolding, ΔH_{cal}, is known to be similar to the van't Hoff enthalpy ΔH_{VH}, which is computed from the dependence of a putative equilibrium constant on temperature under the assumption of two-state behavior. Adapting Eq. (6) of Ref. [19] to the notation used in the present study and making the direct correspondence between model energies and experimental enthalpies, we computed the following quantity:

$$\kappa_2 = \frac{\Delta H_{\mathrm{VH}}}{\Delta H_{\mathrm{cal}}} = \frac{2T_f\sqrt{C_V(T_f)}}{-E^*}. \tag{10}$$

Figure 8(a) shows that the maximal values of κ_2 for the chains of 18 monomers are close to 0.6. Although this value compares well with other simple models, it is significantly below unity and indicative, therefore, of non-two-state pro-

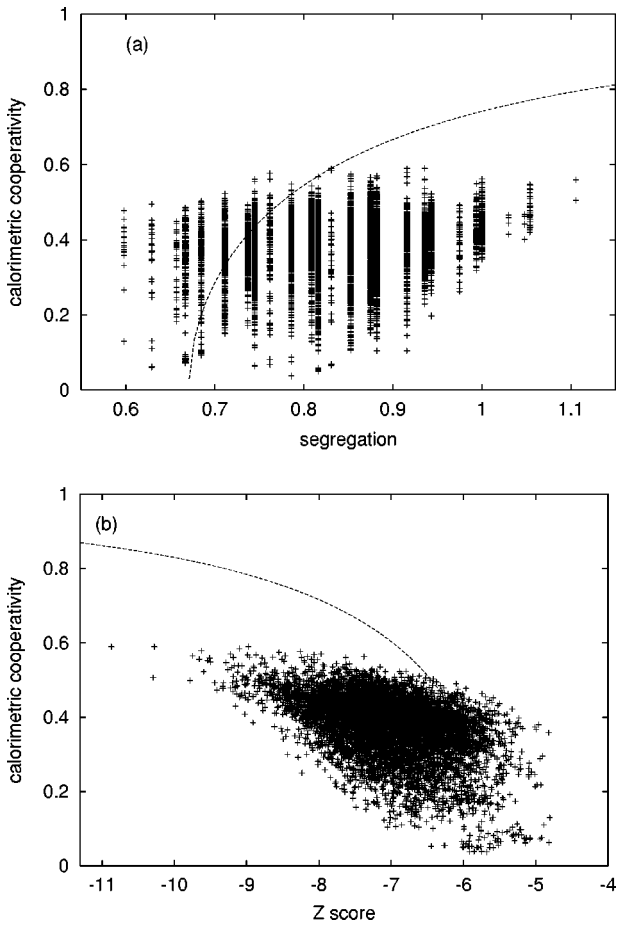

FIG. 8. Relation between folding cooperativity, as measured by Chan's calorimetric parameter κ_2, and structural segregation, as measured by σ (correlation coefficient of 0.162) (a) and relation between κ_2 and Z (correlation coefficient of -0.469) (b). The expected behavior of a closely related cooperativity parameter κ_0 given by Eqs. (11) and (12) are also shown in (b) and (a), respectively, with g_D, the total size of conformational space, taken from column A of Table I and $\sigma_c = 0.51$. All quantities are adimensional. The correlation coefficient between σ and Z for the same set of points is -0.69.

teinlike behavior. Failure to capture this particular aspect of protein folding is likely to result from intrinsic limitations of two-dimensional models in addition to the absence of multibody, explicitly cooperative, interacions [16,19,20]. The correlation between κ_2 and σ, 0.162, is again low, even though the correlation between κ_2 and Z, -0.469, [Fig. 8(b)] is slightly stronger than the correlation between Ω_c and Z.

Mild correlations between cooperativity and segregation and between cooperativity and Z score are clearly consistent with the widely scattered distributions of points shown in Figs. 7 and 8. It should also be noted, however, that there is no reason to expect the dependence of cooperativity on Z and, as a consequence, on σ to be linear. It is actually interesting to compare the observed dependence of κ_2 on σ and Z with the predicted behavior of a closely related cooperativity parameter κ_0 under the assumption of the random energy model for the density of states of the system [16]. From Eq. (22) of Ref. [16] it follows that the dependence the calorimetric cooperativity parameter on Z should be

$$\kappa_2 \approx \kappa_0 = \sqrt{1 - \frac{2 \ln g_D}{Z^2}} \tag{11}$$

which, in combination with Eq. (8), gives the expected dependence on σ

$$\kappa_2 \approx \kappa_0 = \sqrt{1 - \frac{2\sigma_c^2 \ln g_D}{N\sigma^2}}, \tag{12}$$

where g_D is the total size of conformational space or, from column A of Table I, 5 808 335 for $N = 18$. The above expressions are also plotted in Fig. 8 and show that they can, to some extent, predict the values of Z and σ for which successful design is possible and the general trend of κ_2 to increase with the absolute value of these parameters. It is clear, however, that the actual cooperativity tends to be significantly smaller than the estimated value.

III. CONCLUSION

We have investigated the exact dependence of Z scores on structural segregation, as measured by σ, and chain compaction, as measured by C/N, for chains of up to 18 monomers in the square lattice with an unspecific hydrophobic energy function. Sequences were unambiguously designed as N-dimensional vectors in the plain defined by the contact vector representing the target native conformation and the main diagonal. We have found that Z displays a general tendency to decrease linearly with $\sigma\sqrt{N}$, as has been theoretically predicted and previously corroborated by Monte Carlo simulations. Fluctuations around the predicted dependence might be significant, however, particularly at intermediate segregations, and are not completely unexpected since the theoretical analysis was based on simplifying assumptions regarding the unfolded state ensemble. The general behavior of Z is reflected in a dramatic increase in the probability of successful design as σ becomes large, actually approaching unity for maximally segregated conformations, while the probability of successful design as a function of compaction is always much lower. Since the number of maximally segregated conformations is small, however, there is a huge number of successfully designed sequences that correspond to conformations with submaximal values of σ and that cannot be separated from their unsuccessfully designed counterparts by a segregation criterion. Among successfully designed sequences, segregation displays a weak correlation with folding quality, as measured by simple parameters computable from heat capacity curves. This last result appears to arise not only because the correlation between segregation and Z is not perfect, but mainly because the correlation between cooperativity and Z itself is only mild for this set of sequence-structure pairs.

ACKNOWLEDGMENT

M.A.A.B. and A.F.P.A. acknowledge the Brazilian Government agency CNPQ for support.

[1] A.F. Pereira de Araújo, Proc. Natl. Acad. Sci. U.S.A. **96**, 12 482 (1999).
[2] A.F. Pereira de Araújo, J. Chem. Phys. **114**, 570 (2001).
[3] L.G. Garcia, W.L. Treptow, and A.F. Pereira de Araújo, Phys. Rev. E **64**, 011912 (2001).
[4] W.L. Treptow, M.A.A. Barbosa, L.G. Garcia, and A.F. Pereira de Araújo, Proteins: Struct., Funct., Genet. **49**, 167 (2002).
[5] H. Li, R. Helling, C. Tang, and N.S. Wingreen, Science **273**, 666 (1996).
[6] H. Li, C. Tang, and N.S. Wingreen, Proc. Natl. Acad. Sci. U.S.A. **95**, 4987 (1998).
[7] M. Skorobogatiy, H. Guo, and M.J. Zuckermann, Macromolecules **30**, 3403 (1997).
[8] M. Skorobogatiy and G. Tiana, Phys. Rev. E **58**, 3572 (1998).
[9] K. Yue, K.M. Fiebig, P.D. Thomas, H.S. Chan, E.I. Shakhnovich, and K.A. Dill, Proc. Natl. Acad. Sci. U.S.A. **92**, 325 (1995).
[10] E.I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).
[11] R. Melin, H. Li, N.S. Wingreen, and C. Tang, J. Chem. Phys. **110**, 1252 (1999).
[12] R. da Silva, M. da Silva, and A. Caliri, J. Chem. Phys. **114**, 4235 (2001).
[13] A. Kabakçioĝlu, I. Kanter, M. Vendruscolo, and E. Domany, Phys. Rev. E **65**, 041904 (2002).
[14] H.S. Chan and K.A. Dill, Macromolecules **22**, 4559 (1989).
[15] M. Vendruscolo, B. Subramanian, I. Kanter, E. Domany, and J. Lebowitz, Phys. Rev. E **59**, 977 (1999).
[16] H.S. Chan, Proteins **40**, 543 (2000).
[17] H.S. Chan and K.A. Dill, Proteins **24**, 335 (1996).
[18] D. Klimov and D. Thirumalai, Folding Des. **3**, 127 (1998).
[19] H. Kaya and H.S. Chan, Proteins **40**, 637 (2000).
[20] H. Kaya and H.S. Chan, Phys. Rev. Lett. **85**, 4823 (2000).