Identification of amino acid sequences with good folding properties in an off-lattice model

Anders Irbäck,* Carsten Peterson,† and Frank Potthast‡
Complex Systems Group, Department of Theoretical Physics, University of Lund, Sölvegatan 14A, S-223 62 Lund, Sweden

(Received 13 May 1996; revised manuscript received 23 August 1996)

Folding properties of a two-dimensional toy protein model containing only two amino acid types, hydrophobic and hydrophilic, respectively, are analyzed. An efficient Monte Carlo procedure is employed to ensure that the ground states are found. The thermodynamic properties are found to be strongly sequence dependent in contrast to the kinetic ones. Hence, criteria for good folders are defined entirely in terms of thermodynamic fluctuations. With these criteria sequence patterns that fold well are isolated. For 300 chains with 20 randomly chosen binary residues approximately 10% meet these criteria. Also, an analysis is performed by means of statistical and artificial neural network methods from which it is concluded that the folding properties can be predicted to a certain degree given the binary numbers characterizing the sequences. [S1063-651X(97)11901-8]

PACS number(s): 87.15.−v, 87.10.+e

I. INTRODUCTION

The protein folding problem is not merely an engineering task — given sequences of amino acid residues compute its three-dimensional (3D) structure by minimizing an appropriately chosen energy function. Since for such models the energy landscape is often rugged, the resulting 3D configurations may be hard to reach and furthermore may not be thermodynamically stable. It has therefore been argued that only those sequences with ''nice'' energy landscapes have survived the evolution [1].

A proper understanding of the thermodynamics and kinetics of protein folding requires studies of simplified toy models where the conditions can be somewhat controlled. For the choice of such models two major pathways exist. The currently most popular choice is lattice models with contact term interactions; see, e.g., Refs. [1–4]. This approach has the advantage that the ground states are known, but at the same time it has the potential danger that the energy landscape contains artifacts from the discrete description of space. Alternatively, one may use a continuum model with simplified interactions (see, e.g., Refs. [5–7]), in which case substantial simulations are needed to map out the ground states. On the other hand, in this case properties of the energy landscape should be closer to those of the real world.

The aim of this paper is twofold — to map out the folding properties of the two-dimensional continuum model of [8,7], hereafter denoted the *AB* model, and to analyze how the folding properties depend upon the sequences using statistical and state-of-the-art regression methods.

The folding properties of the *AB* model are investigated with respect to thermodynamics and kinetics given a set of thoroughly simulated sequences. In total 300 sequences of 20 hydrophobic and hydrophilic residues ($+1$ and -1, respectively) are studied using an efficient dynamical-parameter algorithm (see Ref. [7] and references therein). The thermodynamic properties are studied using the mean-square distance δ^2 between different configurations. A low average value $\langle \delta^2 \rangle$ signals that the chain exists in a state with well-defined structure. It turns out that $\langle \delta^2 \rangle$ exhibits very strong sequence dependence in contrast to the kinetic properties. Based on this we formulate criteria for good folders entirely based on the distribution of δ^2. Using these criteria roughly 10% of the 300 generated and studied sequences survive as good folders.

Next we pose the question of what characterizes the good folders in terms of sequence patterns. Rather than analyzing the ''bare'' binary sequences of hydrophobicity, we focus on effective variables such as random walk representations, block fluctuations, and the number of $+1$ embedded between two -1. This has the virtue that the analysis will capture long range correlations in addition to the local ones. We investigate how $\langle \delta^2 \rangle$ depends upon these quantities. This is done using tools of varying sophistication — covariance matrix and feedforward artificial neural networks (ANN). Using ANN we predict $\langle \delta^2 \rangle$ given the sequence. With our limited data set the results look very promising. Indeed, the folding properties strongly depend upon sequence patterns. These findings give further evidence of the nonrandomness reported in Ref. [10].

Hydrophobicity is widely believed to play a central role in the formation of 3D protein structures. In Ref. [10] the question of whether proteins originate from random sequences of amino acids was addressed by means of a statistical analysis in terms of blocked and random walk values formed by binary hydrophobic assignments of the amino acids along the protein chains. The results, which were based upon proteins in the SWISS-PROT data base [11], convincingly demonstrated that the amino acid sequences in proteins differ from what is expected from random sequences in a statistical significant way. In Ref. [10] also preliminary results from the *AB* model using the same data as in this work were subject to the same statistical analysis. The interesting observation was made that the *AB* model sequences that fold well according to low $\langle \delta^2 \rangle$-value criteria exhibit similar deviations from randomness as for the functional proteins. The deviations from randomness can be interpreted as originating from

*Electronic address: irback@thep.lu.se

†Electronic address: carsten@thep.lu.se

‡Electronic address: frank@thep.lu.se

anticorrelations in terms of an Ising spin model for the hydrophobicities.

Our studies of the *AB* model are limited to two dimensions in order to be able to analyze many sequences within limited CPU resources. How realistic this approximation is can of course be questioned. The system may be ''stiffer'' than a three-dimensional one when it comes to swapping monomer positions.

This paper is organized as follows. In Sec. II we briefly describe the model and generation of sequences. The Monte Carlo method and what is being measured can be found in Sec. III. The thermodynamics and kinetics of the system are described in Sec. IV, whereas Sec. V contains our statistical and ANN analysis. In Sec. VI we briefly review the results from Ref. [10] comparing deviations from nonrandomness in the two-dimensional *AB* model with those of functional proteins. A brief summary can be found in Sec. VII.

II. THE MODEL

A. General formulation

The *AB* model consists of two kinds of monomers, *A* and *B*, respectively. These are linked by rigid bonds of unit length to form linear chains living in two dimensions. For an *N*-mer the sequence of monomers is described by the binary variables $\sigma_1, \ldots, \sigma_N$ and the configuration by the angles $\theta_2, \ldots, \theta_{N-1}$, where θ_i denotes the bend angle at site *i* and is taken to satisfy $|\theta_i| \leq \pi$. The energy function is given by

$$E(\theta,\sigma)=\sum_{i=2}^{N-1} E_1(\theta_i)+\sum_{i=1}^{N-2}\sum_{j=i+2}^{N} E_2(r_{ij},\sigma_i,\sigma_j), \quad (1)$$

where

$$E_1(\theta_i)=\tfrac{1}{4}(1-\cos\theta_i),$$

$$E_2(r_{ij},\sigma_i,\sigma_j)=4[r_{ij}^{-12}-C(\sigma_i,\sigma_j)r_{ij}^{-6}] \quad (2)$$

and $r_{ij}=r_{ij}(\theta_{i+1}, \ldots, \theta_{j-1})$ denote the distances between sites *i* and *j*. The term $E_1(\theta_i)$ favors alignment of three successive sites; $i-1$, *i*, and $i+1$. The nonbonded interactions E_2 are Lennard-Jones potentials with a species-dependent coefficient $C(\sigma_i,\sigma_j)$, which is taken to be 1 for an *AA* pair (strong attraction), 1/2 for a *BB* pair (weak attraction), and $-1/2$ for an *AB* pair (repulsion). Consequently, there is an energetic preference for separation between the two kinds of monomers. In fact, it was demonstrated in [8] that ground-state configurations tend to have a core consisting mainly of *A* monomers, which shows that *A* and *B* monomers behave as hydrophobic and polar residues, respectively. The behavior of the model at finite temperature *T* is defined by the partition function

$$Z(T,\sigma)=\int\left[\prod_{i=2}^{N-1} d\theta_i\right]\exp[-E(\theta,\sigma)/T]. \quad (3)$$

B. The sequences

In total 300 sequences were drawn randomly from the set of all distinguishable chains with 14 *A* and 6 *B* monomers. Our motivation for this somewhat arbitrary choice of *A*/*B* ratio is that there are thermodynamically stable structures at relatively high temperatures for this ratio [7]. This set contains 19 980 sequences, whereas the total number of sequences with the same composition is 38 760. Among the 300 sequences 4 are symmetric. The 300 distinguishable sequences can be taken as 300 independent sequences drawn from the distribution of all sequences with double weight for every asymmetric sequence.

III. SIMULATIONS

A. Methods

We have performed numerical simulations of both the thermodynamic and kinetic behavior of the 300 randomly selected sequences. At low temperature the system is in a folded phase with high free-energy barriers, which makes conventional simulation methods very time consuming. As in Ref. [7], we therefore employ the dynamical-parameter method for the thermodynamic simulations. In this approach one tries to accelerate the simulation by letting some parameter of the model become a dynamical variable, which takes values ranging over a definite set. In the present work we have taken the temperature as a dynamical parameter (''simulated tempering'' [9]), which means that we simulate the joint probability distribution

$$P(\theta,k)\propto\exp[-g_k-E(\theta,\sigma)/T_k], \quad (4)$$

where T_k, $k=1, \ldots, K$, are the allowed values of the temperature. The g_k's are tunable parameters, which must be chosen carefully for each sequence. The determination of these parameters has been carried out by the same methods as in Ref. [7].

The joint distribution $P(\theta,k)$ is simulated by using an ordinary Metropolis step [12] in *k* and a hybrid Monte Carlo update [13] of θ. The hybrid Monte Carlo update is based on the evolution arising from the fictitious Hamiltonian

$$H_{\mathrm{MC}}(\pi,\theta)=\frac{1}{2}\sum_{i=2}^{N-1}\pi_i^2+E(\theta,\sigma)/T, \quad (5)$$

where π_i is an auxiliary momentum variable conjugate to θ_i. The first step in the update is to generate a new set of momenta π_i from the equilibrium distribution $P(\pi_i)\propto\exp(-\pi_i^2/2)$. Starting from these momenta and the old configuration, the system is evolved through a finite-step approximation of the equations of motion. The configuration generated in such a trajectory is finally subject to an accept-or-reject question, which removes errors due to the discretization of the equations of motion. The hybrid Monte Carlo update has two tunable parameters, the step size ϵ and the number of steps *n* in each trajectory.

The dynamical-parameter method greatly improves the frequency of transitions between different free-energy valleys, as compared to plain hybrid Monte Carlo method. In Ref. [7] a speed up factor of almost 10^3 was observed for system size $N=10$. For $N=20$ we expect the gain to be even larger.

In our simulations we used a set of $K=13$ allowed temperature values, which were equidistant in $1/T$ and ranging from 0.15 to 0.60. Each hybrid Monte Carlo trajectory was

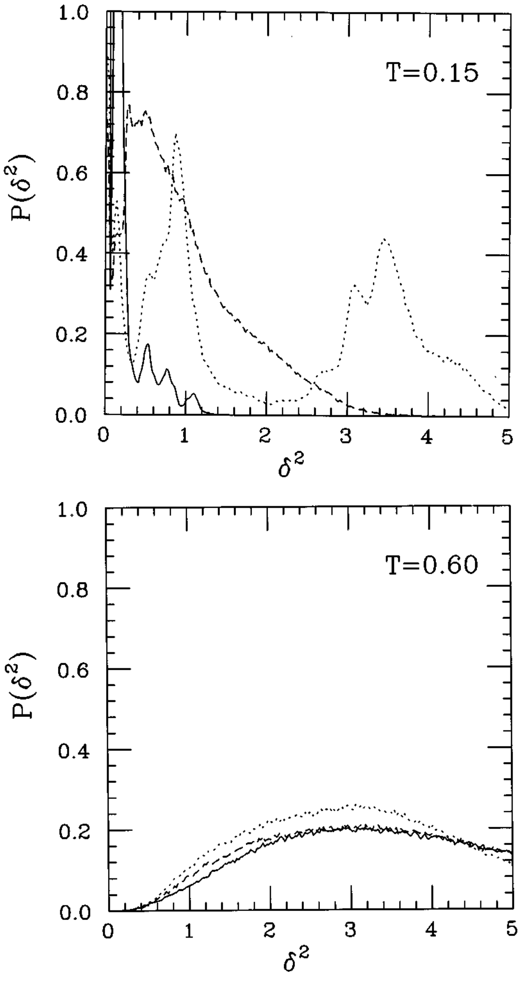

FIG. 1. $P(\delta^2)$ for $T=0.15$ and $T=0.60$, respectively, for the sequences in Table I; 81 (solid line), 10 (dashed line), and 50 (dotted line). For $T=0.15$ the distribution for sequence 81 is dominated by two narrow peaks at small δ^2, which extend outside the figure with a maximum value of around 20.

followed by one Metropolis step in k. The step size parameter ϵ was taken to vary with T, from 0.0025 at $T=0.15$ to 0.005 at $T=0.6$, while $n=100$ was held fixed. For the typical sequence the average acceptance rate was around 95% for the θ update and 65% for the k update. For each sequence a total of 440 000 update cycles were carried out, which requires around 4 CPU hours on a DEC Alpha 2000.

In order to study the kinetic behavior of the model, we have performed hybrid Monte Carlo simulations at different fixed values of T. Starting from random coils, we study the rate of the subsequent relaxation process. While our hybrid Monte Carlo dynamics is certainly different from any real dynamics, it is still a small-step evolution, so the system has to pass through the free-energy barriers. Hence, we expect relaxation times obtained in this way to reflect the actual kinetic properties of the system.

Our simulations of the kinetics have been performed for five different T. The trajectory length $n\epsilon=0.25$ was held fixed, and the average acceptance rate was typically 85% or higher.

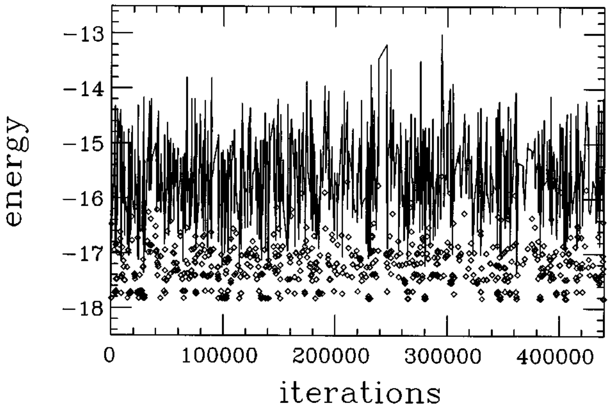

FIG. 2. Evolution of the quenched (diamonds) and unquenched (line) energies in the simulation of sequence 10 (see Table I). Measurements were taken every ten iterations. Shown are the data corresponding to the lowest allowed temperature.

B. Measurements

In our thermodynamic simulations the main goal is to find out whether the system exists in a state with well-defined shape. To address this question, we introduce the usual mean-square distance between configurations. For two configurations a and b we define

$$\delta_{ab}^2=\min\frac{1}{N}\sum_{i=1}^{N}|\overline{x}_i^{(a)}-\overline{x}_i^{(b)}|^2, \tag{6}$$

where $|\overline{x}_i^{(a)}-\overline{x}_i^{(b)}|$ denotes the distance between the sites $\overline{x}_i^{(a)}$ and $\overline{x}_i^{(b)}$ $(\overline{x}_i^{(a)},\overline{x}_i^{(b)}\in R^2)$, and where the minimum is taken over translations, rotations, reflections, and orientations. The probability distribution of δ^2 for fixed temperature T and sequence σ is given by

$$P(\delta^2)=\frac{1}{Z(T,\sigma)^2}\int d\theta^{(a)}d\theta^{(b)}\delta(\delta^2-\delta_{ab}^2)\times e^{-E(\theta^{(a)},\sigma)/T}e^{-E(\theta^{(b)},\sigma)/T}, \tag{7}$$

where $\delta(\cdot)$ denotes the Dirac delta function. $P(\delta^2)$ is a very useful quantity [5] for describing the magnitude of the relevant thermodynamic fluctuations, and can be determined numerically. In Fig. 1 we show three examples of δ^2 distribution at two different temperatures. In what follows we will also frequently use the mean of $P(\delta^2)$,

$$\langle\delta^2\rangle=\int d\delta'^2P(\delta'^2)\delta'^2. \tag{8}$$

The energy level spectrum can be studied by using a quenching procedure, where whenever the lowest allowed temperature value is visited, the system is quenched to zero temperature by means of a conjugate gradient minimization. With this method the ground states are found for most of the sequences. One reason for believing this is that the two or four symmetry-related copies of the lowest-lying minimum were all visited in the simulations. In Fig. 2 we show the evolution

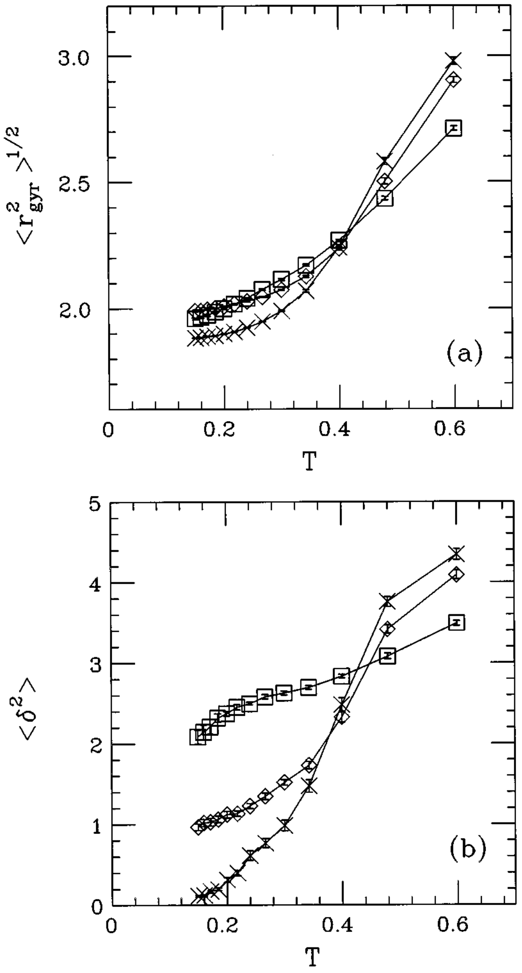

FIG. 3. The temperature dependence of (a) the radius of gyration $\langle r^2_{\mathrm{gyr}}\rangle^{1/2}$ and (b) $\langle\delta^2\rangle$ for the sequences in Table I: 81 (×), 10 (◇), and 50 (□).

of the quenched and unquenched energies in one of the simulations.

IV. THERMODYNAMIC AND KINETIC PROPERTIES

In this section we present the results of our thermodynamic and kinetic simulations. Based on these results, we then formulate criteria for good folding sequences.

A. Thermodynamics

The thermodynamic behavior of the *AB* model has been studied previously for chain lengths $N=8$ and 10 [7]. This study showed that whether or not the chain exhibits a well-defined structure depends strongly on the sequence, at fixed temperature. Using the dynamical-parameter method, we are now able to extend these calculations to $N=20$. The results obtained closely resemble those for $N=8$ and 10 [7]; in particular, they show that the sequence dependence of the thermodynamic behavior remains strong for $N=20$.

In order to illustrate this, we show in Figs. 1 and 3 results for the three sequences in Table I. In Fig. 1 the δ^2 distributions are shown for the lowest and highest temperatures studied, $T=0.15$ and 0.60. At $T=0.60$ the distributions are similar, and show that the fluctuations in shape are large. At $T=0.15$ $P(\delta^2)$ is, by contrast, strongly sequence dependent. We see that one chain exists in a state with very well-defined shape, while the other two still undergo large fluctuations. The differences in $P(\delta^2)$ are clearly reflected in the mean value $\langle\delta^2\rangle$, which is shown as a function of T in Fig. 3(b).

TABLE I. Examples of three sequences.

81	*AAABAAABAAABBAABBAAA*
10	*ABAAABBAABAAAAAAAABB*
50	*BABAAAAAAABAAAABAABB*

Although these three chains behave in very different ways, they have similar extension. In fact, from Fig. 3(a) it can be seen that the differences in radius of gyration are 10% or smaller at all the temperatures studied. Also, we note that the radius of gyration decreases gradually with decreasing temperature. No abrupt changes can be seen.

B. Kinetics

Next we study the time needed by the system to find the minimum energy configuration. Using hybrid Monte Carlo dynamics, we monitor the mean-square distance δ_0^2 to this configuration [see Eq. (6)]. The simulations are started from random coils, and as a criterion of successful folding we use the condition $\delta_0^2<0.3$. At low temperatures the folding time fluctuates widely, which makes the average folding time difficult to measure. For this reason we have chosen to measure the probability of successful folding within a given time [1,15]. Following Ref. [1], this quantity will be called the foldicity. In our simulations the maximum allowed folding time is set to 5000 trajectories. For each of the 300 sequences we studied five different temperatures, $T=0.15, 0.18, 0.24, 0.34$, and 0.60. For each T we carried out 25 simulations for different initial configurations.

The foldicity is expected to be low both at high and low temperature. At low temperature the suppression is due to the ruggedness of the free-energy landscape. At high temperature folding is slow because the search is random.

For a majority of the sequences studied we find that the foldicity exhibits a peak in the interval $0.15\leqslant T\leqslant 0.60$. In order to get precise estimates of the height and location of the peak, one would clearly need more data points. However, the available data demonstrate that the position of the peak is fairly sequence independent. In fact, for 243 of the 300 sequences we obtained a higher foldicity at $T=0.34$ than at the other four temperatures. Also, we note that all sequences have a foldicity of 12% or higher at $T=0.34$, as can be seen from Fig. 4. Therefore, it appears that the kinetic behavior has a relatively weak sequence dependence.

In order to illustrate the implications of this, we have plotted in Fig. 5 the foldicity in two different ways. In Fig. 5(a) foldicity is plotted against temperature, and in Fig. 5(b) it is plotted against $\langle\delta^2\rangle$. Again, we use the three sequences in Table I as examples. From (a) it can be seen that, at a given temperature, the foldicity is roughly similar for the three sequences. Nevertheless, it follows from (b) that their folding properties are very different; sequence 10 has a well-defined shape at the point where the system freezes, which is not true for the other two sequences.

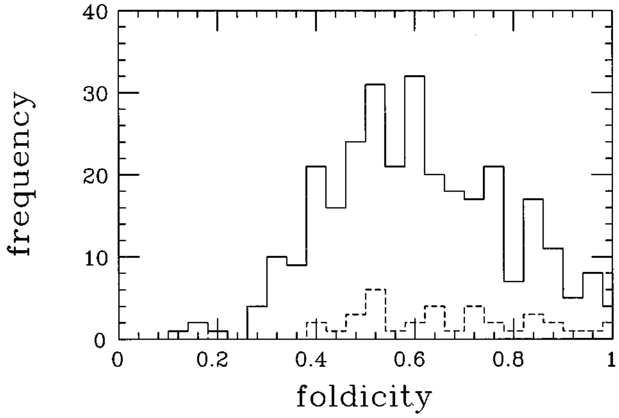

FIG. 4. Histogram of the foldicity at $T=0.34$ for the full set of 300 sequences (solid line) and for a subset of 37 sequences that satisfy the thermodynamic stability condition in Eq. (9) (dashed line).

C. Folding criteria

A good folder is a sequence that , at some value of the parameter T, exists in a unique and kinetically accessible state with well-defined shape. One way to find out whether or not a given sequence meets this requirement is to first locate the lowest temperature at which folding is fast, $\tilde{T}$, and then study the thermodynamic behavior at this temperature. If the sequence is found to exhibit a well-defined structure at $\tilde{T}$, then it is a good folder. Alternatively, this can be formulated in terms of the folding temperature T_f, defined as the temperature where the dominance of a single state sets in; a good folder is a sequence with $T_f > \tilde{T}$.

In our determination of good folders we take the kinetic quantity $\tilde{T}$ to be same for all sequences. This means that our classification is entirely determined by the thermodynamic behavior at a fixed temperature, which we take as $\tilde{T}=0.15$ [cf. Fig. 5(a)]. Our motivation for this simplifying approximation is that the kinetic behavior has a relatively weak sequence dependence, as discussed in the previous subsection.

With this approximation, a natural criterion for good folders would be to require that $\langle \delta^2 \rangle < \tilde{\delta}^2$ for some $\tilde{\delta}$. Such a cut is appropriate for most sequences. However, some care has to be exercised since one might encounter situations where $P(\delta^2)$ has a tiny but distant outlier bump, which can make $\langle \delta^2 \rangle$ large even though the system spends a large fraction of the time very near one particular configuration. Taking this into account we define a sequence to be a good folder if

$$\langle \delta^2 \rangle < \tilde{\delta}^2 \quad \text{or} \quad P(\delta^2 < 0.1) = \int_0^{0.1} d\delta'^2 P(\delta'^2) > \tilde{P} \qquad (9)$$

with $\tilde{\delta}^2 = 0.3$ and $\tilde{P} = 0.35$.

With this choice of parameters, we find that 24 of our sequences satisfy $\langle \delta^2 \rangle < \tilde{\delta}^2$ whereas 30 satisfy $P(\delta^2 < 0.1) > \tilde{P}$. Our set of good folders, which satisfy one or both of these conditions, contains 37 of the 300 sequences, or 12%.

The precise number of sequences classified as good folders depends, of course, on the choice of $\tilde{\delta}$ and $\tilde{P}$, which to some extent is arbitrary. However, it should be stressed that for most of the sequences the classification is unambigiuous in the sense that it is insensitive to small changes of $\tilde{\delta}$ and $\tilde{P}$. Furthermore, we note that the foldicity distribution for good folders is similar to that for all sequences, as can be seen from Fig. 4. If these distributions had been different, the use of a sequence independent $\tilde{T}$ would have been unjustified.

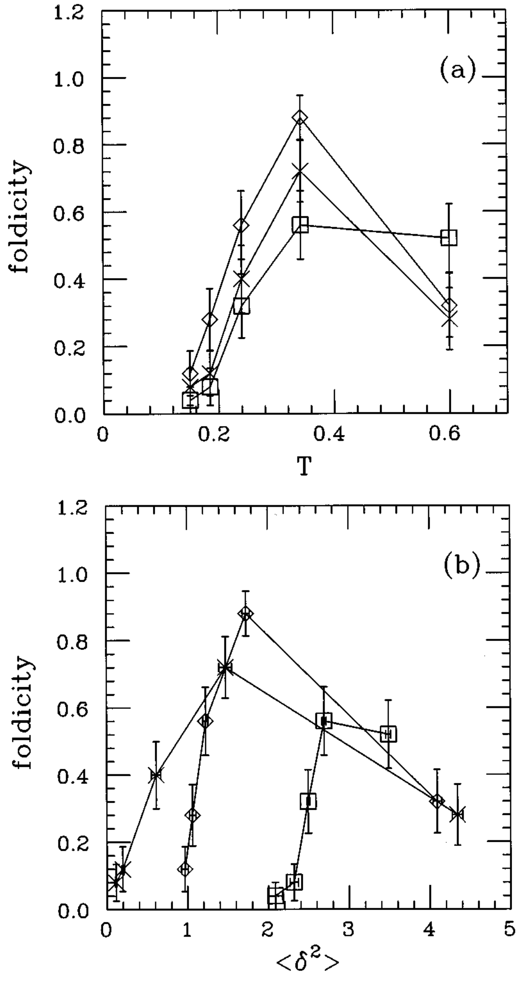

FIG. 5. Foldicity against (a) T and (b) $\langle \delta^2 \rangle$ for the sequences 81 (×), 10 (◇), and 50 (□); see Table I.

V. SEQUENCE CHARACTERISTICS OF GOOD FOLDERS

In this section we analyze how $\langle \delta^2 \rangle$ depends upon the binary patterns of the sequences. This is done in two steps. First we make a statistical analysis in terms of correlations. Second, we employ a feedforward ANN to predict $\langle \delta^2 \rangle$ given the sequence as input.

A. Choice of variables

It turns out to be enlightening and profitable to transform the original binary patterns into more global variables prior to performing the statistical and ANN analysis. The following variables are formed:

Random walk representations — r_n. In order to build in some long range correlation properties we consider random walk representations

$$r_n=\sum_{i=1}^{n}\sigma_i, \quad n=1,\ldots,N \tag{10}$$

Block fluctuations—$\psi_i^{(s)}$ and $\psi^{(s)}$. For a block size s we define the variables [10]

$$\sigma_i^{(s)}=\sum_{j=1}^{s}\sigma_{(i-1)s+j}=r_{is}-r_{(i-1)s}, \quad i=1,\ldots,N/s \tag{11}$$

In order to efficiently capture the fluctuations of the block variables we introduce the normalized variables

$$\psi_i^{(s)}=\frac{1}{K}\left(\sigma_i^{(s)}-\frac{s}{N}\sum_{j=1}^{N/s}\sigma_j^{(s)}\right)^2, \quad i=1,\ldots,N/s \tag{12}$$

and the (normalized) mean-square fluctuation of the block variables

$$\psi^{(s)}=\frac{s}{N}\sum_{i=1}^{N/s}\psi_i^{(s)}, \tag{13}$$

where the constant K can be found in Ref. [10].

Number of internal A's—N_{IA}. This is the number of $\sigma=+1$ surrounded on both sides by $\sigma=+1$. For boundary residues, a single adjacent $\sigma=+1$ is sufficient for giving a count. The reason for this choice of variable is that in the homopolymer limit, $AAAA\cdots A$, the energy landscape is degenerate and, hence, the fluctuations are large. Therefore, one expects long stretches of A's (or B's) to be rare in good folders, and that N_{IA} tends to be low for such sequences.

Number of clumps—N_C. This quantity is defined as the number of clumps of $\sigma=\pm1$. The reason for including this variable is similar to what was argued for N_{IA} above; it seems natural to expect a high N_C for good folders.

We expect these preprocessed variables, which of course are not independent of each other, to shed more light on the structure than the ''raw'' $\sigma=\pm1$ ones, when it comes to relate the sequences to $\langle\delta^2\rangle$.

In this section we will make use of all patterns, even those generated by symmetry giving rise to the same $\langle\delta^2\rangle$. As mentioned in Sec. II B, 4 of the 300 original sequences are symmetric, which means that we have a total of 596 patterns at our disposal.

B. Correlations

To what extent is $\langle\delta^2\rangle$ correlated with r_n, $\psi_i^{(4)}$, $\psi^{(4)}$, N_{IA}, and N_C? In Fig. 6 the correlation between $\langle\delta^2\rangle$ and these variables is shown.

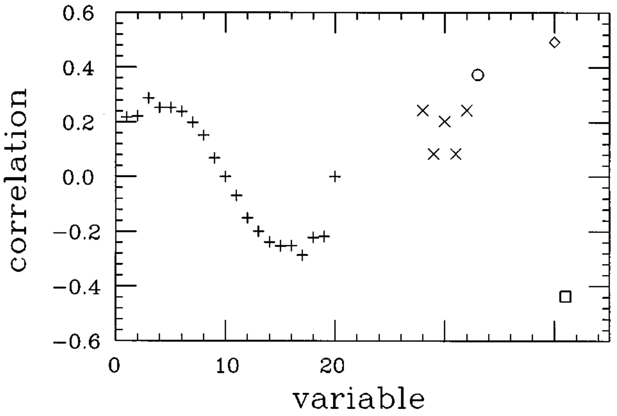

FIG. 6. Correlations of $\langle\delta^2\rangle$ against r_n [$n=1,\ldots,20$] (+), $\psi_i^{(4)}$ [$i=1,\ldots,5$] (×), $\psi^{(4)}$ (○), N_{IA} (◇), and N_C (□).

As can be seen from Fig. 6 the correlations are substantial. It should also be mentioned that the same correlation patterns emerge when reducing the data set by a factor of 4. The symmetry for r_n around $n=10$ is inherent from the way we have generated the data, and leaves us with nine independent measurements, corresponding to, e.g., r_1 through r_9. The following general observations can be made: (1) $\langle\delta^2\rangle$ and r_n exhibit sizable positive correlations near the end points, implying that proteins with many A's near the ends do not fold well. (2) $\langle\delta^2\rangle$ and the block fluctuations, $\psi_i^{(4)}$ and $\psi^{(4)}$, also show strong positive correlations. This is consistent with what was observed for real proteins with limited net hydrophobicity in Ref. [10], where these variables are anticorrelated as compared to what is expected from random hydrophobicity distributions. (3) The strongest positive correlation is between $\langle\delta^2\rangle$ and the number of internal A's (N_{IA}). This is in line with what is expected according to the motivation when introducing N_{IA} above. (4) Related to positive correlation of N_{IA} is the strong anticorrelation between $\langle\delta^2\rangle$ and the number of clumps N_C.

C. Artificial neural networks

Given the substantial correlations between $\langle\delta^2\rangle$ and the various quantities formed out of the sequence patterns, it should be possible to make a regression model. If enough data are available for an efficient and reliable fit one should be able to predict the folding properties given a binary sequence. The state-of-the-art technique for such modeling are feedforward ANN (see, e.g., Ref. [14]), which will be used here. This method has the advantage of capturing nonlinear dependencies in a generic way in contrast to standard linear regression approaches. In our case the feedforward ANN consists of an input layer representing the variables defined in Sec. V A above, an output unit for $\langle\delta^2\rangle$ and a set of hidden units in order to model nonlinearities. The weights connecting the nodes are the parameters of the system. In order to avoid overfitting, the number of weights (parameters) should be less than the number of ''training'' patterns. Also, some of the patterns should be set aside for ''testing.'' Using all the quantities (only r_1 through r_9 due to symmetries) defined in Sec. V A with 5 hidden units implies $17\cdot5+5=90$ parameters.

The network was trained using the JETNET 3.0 package [16] with an initial learning rate $\eta_0=0.5$, which decreases according to $\eta_k=0.998\eta_{k-1}$ and momentum $\alpha=0.7$. In order to obtain as reliable performance as possible, the K-fold cross validation procedure was used, where the data set was randomly divided into K equal parts. Each of the K different parts was once used as a test set, while the remaining $K-1$ sets were used for training. In our problem sets with $K=2$, 3, and 4 were used. In Fig. 7 the resulting prediction, $\langle\hat{\delta}^2\rangle$, is compared with the true values, $\langle\delta^2\rangle$, for a

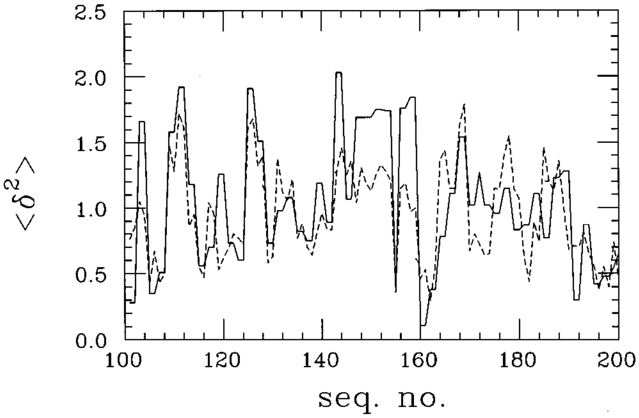

FIG. 7. Comparison of predicted and true values, $\langle\hat{\delta}^2\rangle$ and $\langle\delta^2\rangle$, respectively, using $K=3$ for sequences 100 – 200. In order to guide the eye, lines connecting the points are drawn. The full line represents $\langle\delta^2\rangle$ and the dashed line $\langle\hat{\delta}^2\rangle$.

representative subset of the sequences using $K=3$ is shown. As can be seen from Fig. 7 the predicted values follow the true values pretty well even though sometimes quantitative agreement is lacking. The degree of success can be quantified using the standard error of prediction,

$$R=\frac{1}{\sigma^2}\frac{1}{N}\sum\left(\langle\hat{\delta}^2\rangle-\langle\delta^2\rangle\right)^2, \tag{14}$$

where σ is the standard deviation of $\langle\delta^2\rangle$ for all 596 patterns. For $K=3$ and 4 we find $R=0.69\pm0.02$. This value deteriorates somewhat if a smaller subset of the data is used, which indicates that with a larger available data set the performance is likely to improve. In the long term one would like to predict a classification (folder or nonfolder) but the limited folding data at our disposal do not yet allow for that.

VI. NONRANDOMNESS AND FOLDING

In Sec. V we studied the difference between folding and nonfolding sequences in the AB model. This could be done in a controlled way due to the fact that we have unbiased samples of folding and nonfolding sequences at our disposal. To assess the applicability of our methods to real amino acid sequences is more difficult, and we shall not deal here with the problem of predicting the behavior of individual amino acid sequences. However, we would like to stress that the binary hydrophobicity patterns corresponding to a large class of functional proteins do exhibit interesting similarities with folding sequences in the AB model, as was demonstrated in Ref. [10]. The proteins sequences considered in Ref. [10] have a limited net hydrophobicity

$$X=\frac{N_+-Np}{\sqrt{Np(1-p)}}, \tag{15}$$

where N_+ is the number of hydrophobic residues, N is the total number of residues, and p is the average of N_+/N over all sequences. In the following, we consider sequences with $|X|<0.5$.

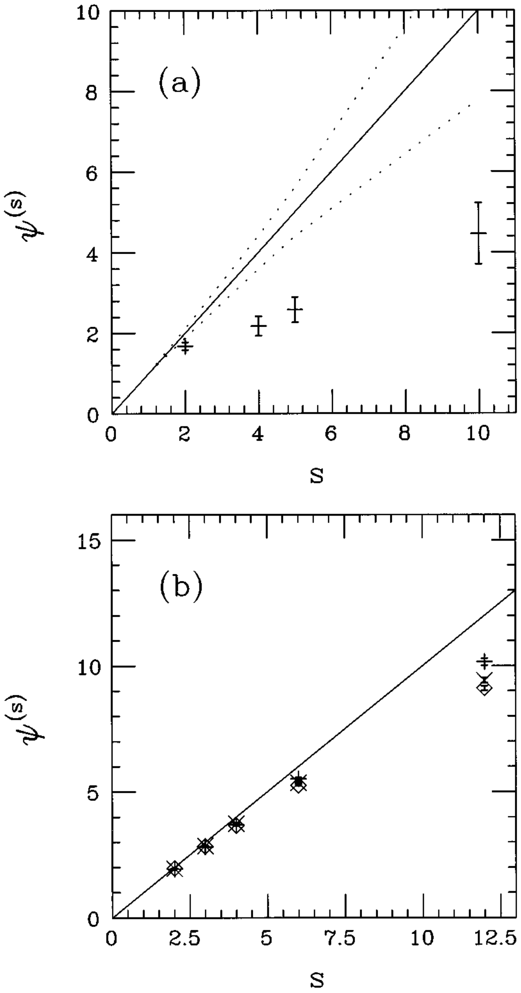

FIG. 8. Mean-square fluctuation of the block variables, $\psi^{(s)}$, against block size s. (a) Good folding sequences in the AB model. Also shown are the mean s (full line) and the $s\pm\sigma$ band (bounded by dotted lines) for random sequences [10]. (b) Functional proteins for $|X|<0.5$, $50<N\leqslant150$ (+, 2457 qualifying proteins), $150<N\leqslant250$ (×, 2228), and $250<N\leqslant350$ (◇, 1642). All data are from the SWISS-PROT data base [11]. The straight line is the result for random sequences [10].

In order to examine these two groups of binary strings, it is instructive to consider the scaling of the fluctuation variable $\psi^{(s)}$ with block size s [10]. In Fig. 8 we show results for this quantity. For both groups, it can be seen that $\psi^{(s)}$ grows significantly slower with s than for random sequences. This behavior implies that the sequence variables σ_i exhibit anticorrelations. A simple way to see that is to consider the one-dimensional Ising model, where each configuration is given a statistical weight

$$P\propto\exp\left(K\sum_{i=1}^{N-1}\sigma_i\sigma_{i+1}\right) \tag{16}$$

For $K=0$ the σ_i's are completely random, while $K>0$ and $K<0$ correspond to ferromagnetic and antiferromagnetic behavior, respectively. It turns out that one obtains results similar to those seen in Figs. 8(a) and 8(b) for $K=-0.25$.

VII. SUMMARY

Fairly detailed studies of the folding properties of a toy model for proteins, containing effective hydrophobicity interactions only, have been performed. The thermodynamic quantity $\langle \delta^2 \rangle$ exhibits a strong sequence dependence in contrast to foldicity, which measure the kinetic properties. Hence criteria for chains with good folding properties have been devised solely in terms of $\langle \delta^2 \rangle$. With these criteria, approximately 10% of the 300 generated sequences are classified as good folders. These conclusions have been possible due to extensive searches using an efficient dynamical-parameter algorithm, which with very large probability visits the ground states. A similar fraction of good folders was obtained in the lattice model study of Ref. [1], where 30 out of 200 randomly chosen sequences were classified as good folders.

Our conclusion that the thermodynamic properties are more important for the classification of folders or nonfolders than the kinetic ones is in line with the lattice model results of Ref. [4]. These authors introduced a kinetic glass transition temperature, which was found to be nearly sequence independent. Although the kinetic studies of Ref. [4] are more elaborate than ours, it should be stressed, however, that our data provide no justification for introducing this transition temperature, which may indicate a difference between the models.

Using statistical and artificial neural network methods, substantial functional dependencies between sequence patterns and $\langle \delta^2 \rangle$ are revealed. With larger statistics it should be possible given the sequence pattern to predict $\langle \delta^2 \rangle$ within a reasonable confidence level, in other words to predict whether a given sequence folds or not.

Related to this strong correlation is the observed pattern of the nonrandomness for the folders, which show similar qualitative behavior with what is observed for real proteins [10].

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

[2] C.J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. USA **90**, 6369 (1993).

[3] H.S. Chan and K.A. Dill, J. Chem. Phys. **100**, 9238 (1994).

[4] J.D. Bryngelson, J.N. Onuchic, N.D. Socci, and P.G. Wolynes, Proteins **21**, 167 (1995).

[5] G. Iori, E. Marinari, and G. Parisi, J. Phys. A **24**, 5349 (1991).

[6] M. Fukugita, D. Lancaster, and M.G. Mitchard, Proc. Natl. Acad. Sci. USA **90**, 6365 (1992).

[7] A. Irbäck and F. Potthast, J. Chem. Phys. **103**, 10 298 (1995).

[8] F.H. Stillinger, T. Head-Gordon, and C.L. Hirschfeld, Phys. Rev. E **48**, 1469 (1993).

[9] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).

[10] A. Irbäck, C. Peterson, and F. Potthast, Proc. Natl. Acad. Sci. USA **93**, 9533 (1996).

[11] A. Bairoch and B. Boeckmann, Nucleic Acids Res. **22**, 3578 (1994).

[12] N.A. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

[13] S. Duane, A.D. Kennedy, B.J. Pendleton, and D. Roweth, Phys. Lett. B **195**, 216 (1987).

[14] J. Hertz, A. Krogh, and R.G. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, CA, 1991).

[15] N.D. Socci and J. N. Onuchic, J. Chem. Phys. **101**, 1519 (1994).

[16] C. Peterson, T. Rögnvaldsson, and L. Lönnblad, Comput. Phys. Commun. **81**, 1 (1994).