Chain length scaling of protein folding time: Beta sheet structures

ARTICLE *in* THE JOURNAL OF CHEMICAL PHYSICS · JULY 2000

Impact Factor: 2.95 · DOI: 10.1063/1.481915

3 AUTHORS, INCLUDING:

Bengt Herbert Kasemo
Chalmers University of Technology
497 PUBLICATIONS **22,865** CITATIONS

Chain length scaling of protein folding time: Beta sheet structures

K. Dimitrievski and B. Kasemo
Department of Applied Physics, Chalmers University of Technology, S-41296 Göteborg, Sweden

V. P. Zhdanov
Department of Applied Physics, Chalmers University of Technology, S-41296 Göteborg, Sweden and Boreskov Institute of Catalysis, Russian Academy of Sciences, Novosibirsk 630090, Russia

(Received 11 February 2000; accepted 11 April 2000)

We present comprehensive 3D lattice Monte Carlo simulations of the folding kinetics of two-turn antiparallel β sheets. The model employed takes into account isotropic nonspecific interactions as in previous flexible heteropolymer models and also orientation-dependent monomer–monomer interactions, mimicking the formation of hydrogen bonds and chain rigidity. The chain length is varied from $N=15$ to 33. For each chain length, we calculate the fastest folding temperature, $T_{\rm fast}$, folding temperature, $T_{\rm fold}$, and glass-transition temperature, T_g. The time-averaged occupation probability of the native state is found to be nearly independent of N at all temperatures. The dependence of $T_{\rm fast}$ and $T_{\rm fold}$ on N is accordingly relatively weak. The temperature interval where the folding is fast rapidly decreases with increasing N. For the chain lengths chosen, $T_{\rm fold}$ slightly exceeds T_g. The dependence of the folding time τ_f on N is well fitted by using the power law, $\tau_f \propto N^{\lambda}$. The exponent λ is found to depend on temperature and on the distribution of nonspecific interactions in the chain. In particular, $\lambda = 2.7-4.0$ at $T=T_{\rm fast}$ and 5.2 at T slightly below $T_{\rm fold}$. Evaluating τ_f in real units at T near $T_{\rm fold}$ yields physically reasonable results. © *2000 American Institute of Physics.* [S0021-9606(00)50526-1]

I. INTRODUCTION

The unique compact native structure of a protein is encoded within its amino acid sequence, which is sometimes referred to as the ''second genetic code'' (the first genetic code lies in the base sequence of DNA molecules, controlling the amino acid sequence of proteins during synthesis). The understanding of the mechanism(s) of protein folding to the native state is a challenging problem related to theory of evolution of life and also having important practical consequences for biotechnology, including such areas as hormones and biological regulatory agents, biological and chemical catalysts, biosensors, etc. The complexity of this intriguing problem has long been recognized to be intimately connected with the fact (often referred to as the Levinthal paradox) that the unique native conformation of a protein cannot be found by a random search among the astronomically large number of possible conformations (this number increases *exponentially* with increasing chain length). Recent advances in the understanding of protein folding are partly connected with extensive Monte Carlo (MC) simulations based on simple lattice models[1–4] (the all-atom off-lattice models require a large amount of computer time and cannot really be used to simulate the folding kinetics from the beginning to the end). New key findings in this field are as follows.

(1) The Levinthal paradox has been solved (see the discussion in Ref. 1) via the realization that the main number of possible protein conformations is irrelevant for folding, because the folding process in practice is governed by a bias, or steering, toward the native state along the effective energy surface sometimes called[5] the ''folding funnel.'' The ''irrelevant'' conformations are thus never visited.

(2) The transition barrier on the way toward the native state has been found to be broad and not especially high (see the review in Ref. 3).

(3) In numerous simulations (see, e.g., Refs. 6 and 7), ith as been demonstrated that folding often occurs via two phases, including a rapid collapse to one of the compact metastable states, followed by slow reconfiguration of the chain to the native structure via additional metastable states.

(4) Factors facilitating or at least correlating with rapid folding have been identified to be: (i) a pronounced energy gap between the native state and the remaining configurations[8] (this criterion is often operative, but not always),[9,10] (ii) a high value of the folding temperature, $T_{\rm fold}$, compared to the glass transition temperature, T_g,[11] and (iii) close-lying values of the collapse temperature, T_c, and $T_{\rm fold}$[10] (the latter two criteria are in fact similar).

(5) The dependence of the folding time, τ_f, on the chain length, N, is found[12–14] to be described by the power law,

$$\tau_f \propto N^{\lambda}, \tag{1}$$

with the model- and temperature-dependent exponent $\lambda = 3-7$.

In addition, the results of lattice-model simulations have been used to illustrate and/or complement general conclu-

sions following from more phenomenological treatments, based, for example, on the energy-landscape theory.[5]

Despite the advances listed above, the field is still far from mature and many interesting questions remain open. The bulk of available simulations is based on highly simplified lattice models representing a protein as a *perfectly flexible heteropolymer* with isotropic monomer–monomer interactions and no energy cost for creation of kinks. The folding kinetics for more realistic models thus remain to be elucidated.

The first natural step in this direction is to remove the main heteropolymer-model limitation related to the independence of the interaction between monomers on their *orientation*. The direct consequence of the latter approximation is that the predicted native protein structures are usually highly compact (close to spherical or cubic). Although the latter is often the case in reality, the native structure of real proteins frequently also contain α helices and/or β sheets.[15,16] Such low-dimensional fragments of proteins are stabilized primarily by hydrogen bonds, i.e., by strongly anisotropic bonds between oxygen and hydrogen atoms belonging to amino-acid residues that are nearest-neighbor (NN) in space but not NN in the polypeptide chain. Hydrogen bonds are formed only if the orientation of amino-acid residues is favorable for their formation (other interactions between amino-acid residues can also depend on their orientation but not so strongly). To take into account the important role of oriented hydrogen bonds in stabilization of proteins, it is desirable to design relatively simple protein models, making it possible to study the folding kinetics of α helices and β sheets. Efforts to solve this problem are just beginning. We have proposed[17] a general lattice model, incorporating orientation-dependent interactions (ODI) between monomers and also the chain rigidity, and used it to explore[13] the chain-length dependence of the folding time of the simplest β sheets. Folding of helical proteins has recently been simulated[18–21] by employing off-lattice models.

An attractive feature of the ODI model[17] is that it is compatible with the standard heteropolymer model. In this article, we show that a combination of these models opens up new opportunities in Monte Carlo simulations of protein folding. Specifically, we present comprehensive MC simulations of the folding kinetics of β sheets. Compared to our earlier studies,[13,17] we use a more realistic way of stabilization of native β sheet structures and explore in more detail the dependence of τ_f on the chain length and temperature. In addition, we present typical folding trajectories in the energy space and histograms characterizing the folding kinetics.

II. MODEL AND ALGORITHM OF SIMULATIONS

As noted in the introduction, our model is based on the lattice approximation, i.e., a protein is schematically viewed as a linear sequence of N amino acids. The structure of the amino acids is not analyzed explicitly. Instead, the protein is simulated by a chain of monomers, which are constrained (along the chain) to be nearest-neighbors on the 3D cubic lattice (each lattice site can be occupied at most once). The equilibrium structure of a protein is defined by monomer–monomer interactions. We take into account isotropic interactions, ϵ_{ij}^{is}, corresponding to the heteropolymer model,[2] anisotropic interactions mimicking the formation of hydrogen bonds, and chain rigidity.[17] To describe anisotropic interaction, we introduce the unit vector $\mathbf{s}_i$ characterizing the orientation of monomer i (Fig. 1). Physically, this vector shows the orientation of an amino-acid residue with respect to the chain. For this reason, the orientations of $\mathbf{s}_i$ along the peptide bonds formed by monomer i with monomers $i-1$ and $i+1$ are prohibited. In the framework of this model, the protein energy is given by

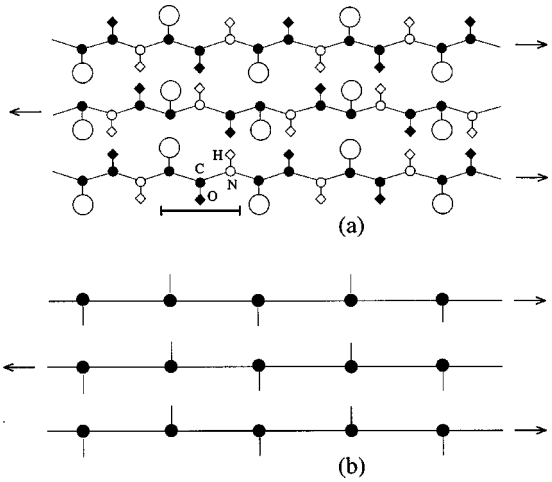

FIG. 1. Ideal antiparallel β sheet. Panel (a) shows the arrangement of carbon (filled circles), nitrogen (small open circles), oxygen (filled diamonds) and hydrogen (open diamonds) atoms, and also amino-acid side chains (large open circles) which are located outside the sheet. The solid bar marks one of the amino-acid residues. Hydrogen bonds are formed between nearest-neighbor hydrogen and oxygen atoms from different rows (note that every residue forms two hydrogen bonds). Panel (b) represents a model arrangement of monomers (filled circles) mimicking amino-acid residues. The orientations of monomers are indicated by thin bars.

$$E_{\mathrm{tot}}=\sum_{|i-j|\geqslant 3}\epsilon_{ij}^{is}\delta(\boldsymbol{r}_{ij}-\boldsymbol{a})+\sum_{|i-j|\geqslant 3}\epsilon_{ij}^{ss}\delta(\boldsymbol{r}_{ij}-\boldsymbol{a})+\sum_{i=1}^{N-1}\mathcal{A}(\mathbf{s}_i\mathbf{s}_{i+1})+\epsilon_k n_k, \tag{2}$$

where r_{ij} is the distance between monomers i and j, a the lattice spacing, $\delta(0)=1$ (zero otherwise), $(\mathbf{s}_i\mathbf{s}_{i+1})$ is a scalar product of the orientation vectors, ϵ_k the energy cost of a kink formation, and n_k the number of kinks.

The first term in the right-hand part of Eq. (2) describes isotropic (e.g., van der Waals) monomer–monomer interactions. The second term corresponds to the hydrogen bonds in topological contacts. These bonds are considered to be formed only provided that (i) one of the monomers is directed toward the other monomer (e.g., $\mathbf{r}_i+\mathbf{s}_i=\mathbf{r}_j$), and (ii) simultaneously the vectors $\mathbf{s}_i$ and $\mathbf{s}_j$ are parallel. If both these conditions are fulfilled, the values of the energy parameters ϵ_{ij}^{ss} (the superscript ss is for the s–s contacts) are negative, i.e., the monomers attract each other (in all the other cases, $\epsilon_{ij}^{ss}=0$). The third term (with $\mathcal{A}>0$), constructed in analogy with the Ising antiferromagnetic spin-spin interaction, describes the orientational dependence of the interaction between NN monomers linked via the peptide bonds. Its goal is

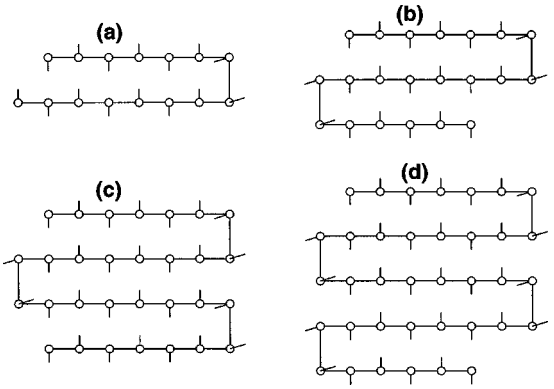

FIG. 2. Examples of (a) one-, (b) two-, (c) three-, and (d) four-turn β sheets formed by B (hydrophilic) monomers.

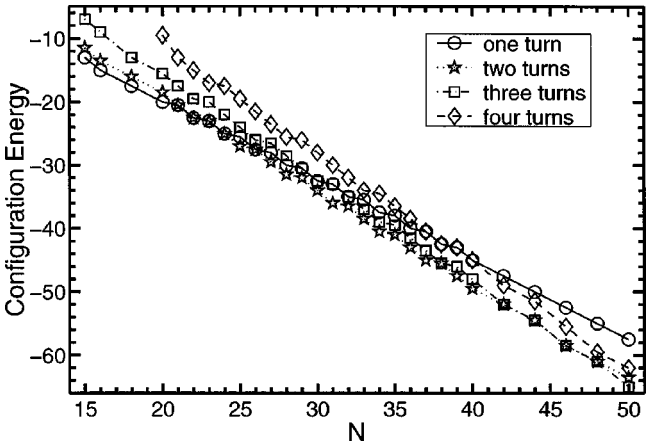

FIG. 3. Conformation energy versus chain length for one-, two-, three-, and four-turn β sheets consisting of B monomers with $\epsilon^{ss}=-1.5$, $\mathcal{A}=0.5$, and $\epsilon_k=1.0$.

to reproduce an antiparallel orientation of NN monomers (Fig. 1). The fourth term takes into account the rigidity of the polypeptide chain.

To mimic the 20-letter code of real proteins, one should introduce scattering of the energies ϵ_{ij}^{is} and ϵ_{ij}^{ss}. The formation of kinks should in general be made monomer-type dependent as well. In our present simulations, we use the simplest two-letter (A and B) code and accordingly have three types of isotropic interactions (ϵ_{AA}, ϵ_{AB}, and ϵ_{BB}). In addition, we consider that all the hydrogen bonds are equivalent (ϵ^{ss} denotes the energy of these bonds). Obviously, our code is poor compared to the 20-letter code of real proteins. However, our ambition is not to cover all the possible situations, but rather to show that this model, with a few steps towards higher sophistication compared to earlier MC simulations, predicts interesting phenomena even when we employ the simplest code. Specifically, our simulations are focused on folding of β sheets. The code used actually turns out to be sufficient for formation of such sheets.

The monomers A and B are assumed to have predominantly hydrophobic and hydrophilic properties, respectively. To reproduce the tendency of hydrophobic monomers to end up inside a water soluble protein (because water–protein interaction favors hydrophilic proteins on the outside), weu se attractive $A-A$ interaction, $\epsilon_{AA}<0$. The $A-B$ interaction is considered to be effectively repulsive, $\epsilon_{AB}>0$, because the nonpolar A monomers are expected to prefer to minimize the number of contacts with the polar B monomers. (ϵ_{AB} could have been made weakly attractive as well as $|\epsilon_{AB}|\ll|\epsilon_{AA}|$.) The interaction ϵ_{BB} is kept zero in order to account for the (supposed) surrounding water, with which the B monomers interact favorably. The formation of hydrogen bonds is energetically favorable, and accordingly we employ $\epsilon^{ss}<0$. The formation of kinks is considered to cost energy, i.e., $\epsilon_k>0$ (this physically reasonable prescription, combined with ODI, results in formation of flat structures).

To simulate the folding kinetics, we have used the standard algorithm for end, corner, and crankshaft monomer moves.[11,17] A monomer is chosen at random. If it is an end monomer, one of the neighboring lattice sites is also selected at random for an end move. If it is not an end bead, then, depending on the position of its neighbors along the chain, it can do either a corner move or a crankshaft move. The possible direction of the latter move is selected at random. The possible orientations of monomers are chosen at random among the allowed orientations. If the move selected violates the excluded volume constraint by moving the monomer to an occupied site, the trial ends. If there is no such constraint, the energies of the original and new configurations are calculated, and the move is realized with the probability given by the Metropolis rule [$W=1$ for $\Delta E\leqslant 0$, and $W=\exp(-\Delta E/k_BT)$ for $\Delta E\geqslant 0$, where ΔE is the energy difference of the final and initial states]. After each attempt of an elementary monomer move, we execute (also by using the Metropolis rule) one attempt to change the orientation of a randomly chosen monomer without changing the monomer positions.

Initially (at $t=0$), the chain is considered to be in the random-coil state. For the time scale, we employ MC steps (MCS) (by definition, one MCS corresponds to N attempts to realize a monomer move).

III. RESULTS OF CALCULATIONS

A. Native structures

All the energetic parameters introduced above are effective, because the solvent is not treated explicitly. Practically, this means that the energetic parameters should be smaller than the typical values ($\sim 3-5$ kcal/mol) of van der Waals or hydrogen bonds. Taking into account that the free energy difference between the denatured and native states of proteins is usually low (often about 5–15 kcal/mol),[16] one can conclude that the values of the model energetic parameters should be lower, about, or slightly larger than k_BT. The bulk of our simulations was executed for $\epsilon^{ss}=-1.5$, $\mathcal{A}=0.5$, and $\epsilon_k=1.0$ (we use dimensionless units). With these physically reasonable parameters, the model easily reproduces β sheets provided that the isotropic monomer–monomer interactions are not too strong.

To understand the type of the native structures predicted by the model, it is instructive first to analyze the chains consisting exclusively of B monomers. In this case (with no isotropic monomer–monomer interactions), the structure of

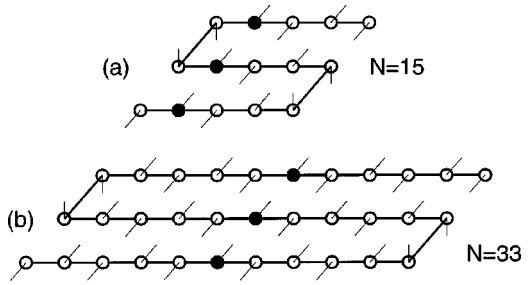

FIG. 4. Arrangements of *A* and *B* monomers (filled and open circles) in the β sheets with $N=15$ (a) and 33 (b). In these cases, ***A*** monomers are, respectively, located on the periphery and in the middle of the sheets. In the text, such sequences are called ***S**equence 1* and ***S**equence 2* irrespective of the chain length.

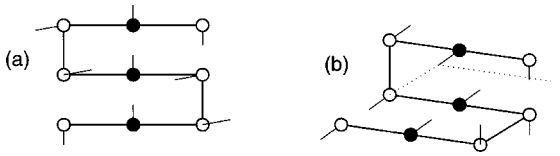

FIG. 5. (a) Target and (b) actual native state of the chain with $N=9$ (filled circles and open circles represent ***A*** and ***B*** monomers, respectively). The energy of conformation (b) is lower than that of conformation (a), because it has an additional hydrogen bond and an additional ''Ising'' contact.

the native conformation depends on the interplay between the energy-demanding formation of kinks and exothermic hydrogen bonds. Kinks are energetically unfavorable, but for long chains their formation can be easily compensated by additional hydrogen bonds. For this reason, the native conformation is intuitively expected to be a sheet with one or more turns (Fig. 2), depending on the chain length. Calculating the energies of such configurations for the chains with $N \leqslant 50$ (see, e.g., the data shown in Fig. 3 for $15 \leqslant N \leqslant 50$), we have found that the one-turn sheet is native for $8 \leqslant N \leqslant 20$ (for $N \leqslant 7$ the native conformation is a straight monomer chain without kinks). For $21 \leqslant N \leqslant 26$, the one- and two-turn sheets have the same energy (except for $N=25$ where the two-turn sheet is lower in energy). For $27 \leqslant N \leqslant 40$, the two-turn sheet is native (except for $N=38$ where the two- and three-turn sheets have the same energy), and then the two- and three-turn sheets have the same energy up to $N=50$, where the three-turn sheet is native.

Introducing ***A*** monomers into the chain makes it possible to stabilize one of the structures described above or change the number of turns in the native state of the chain. Our goal was to form a sequence of two-turn sheets and explore in detail the dependence of the folding kinetics of such sheets on the chain length. Using $\epsilon_{AA}=-3.0$, $\epsilon_{AB}=0.5$, and $\epsilon_{BB}=0$ and simulating the behavior of chains with various distributions of ***A*** monomers, we have found that the desirable native structures can be obtained, e.g., by introducing three ***A*** monomers into the periphery (near the turns) or middle of a sheet as shown in Fig. 4. The fact that these structures, referred to below as ***S**equence 1* and ***S**equence 2*, are really native for $N=15$, 18, 21, 24, 27, 30, and 33 was verified by using MC runs up to 3×10^9 MCS. For shorter chains with $N=9$ and 12, this way of stabilization of the two-turn sheets does not work [see, e.g., Figs. 5(a) and 5(b)]. For longer chains, the formation of native states with more than two turns becomes energetically favorable only for $N>100$.

Changing the distribution of ***A*** monomers (number and position), one can get other desirable native structures. For example, the ideal three-turn sheets can be obtained by employing the distribution shown in Fig. 6 or other similar distributions.

For a given chain length, the energies of ***S**equence 1* and *2* chosen in our present study are equal. Physically, however, these structures are slightly different. In particular, the number of conformations resulting in the formation of bonds between *A* monomers is larger in the latter case, because the distance (along the chain) between all these monomers is considerable. Thus, the folding time for ***S**equences 2* is expected to be slightly shorter. Our results shown below in Figs. 14 and 15 indicate that this is usually the case.

Finally, it is appropriate to note that in our simulations below the conformation is considered to be native if the ''back-bone'' structure coincides with the target β sheet structure, i.e., if the chain has the form of the desirable β sheet regardless of the monomer orientations (fluctuations in the monomer orientations are in fact inevitable, but they practically do not change the protein structure).

B. Folding at different temperatures

The protein folding kinetics are customarily characterized by using the folding time, $\tau_f \equiv \langle t_{\mathrm{fp}} \rangle$, defined as the mean first passage time from unfolded initial states to the native state. The dependence of τ_f on temperature for our model is exhibited in Fig. 7. Specifically, Fig. 7 shows τ_f as a function of T for ***S**equence 1* with $N=15$, 24, and 33 (the results for ***S**equence 2* are similar). As usual, the folding is seen to be relatively fast at intermediate temperatures. At high temperatures, the process is slow because the driving force for folding is weak (the chain rapidly fluctuates between disordered states, randomly searching for the native state). At low temperatures, the folding kinetics are slowed down, because the chain gets trapped in metastable states. At temperatures in between, the conditions for folding are optimal, i.e., the process is effectively driven by the shape of the energy landscape. The temperature interval where the folding is fast rapidly decreases with increasing ***N***, since larger ***N*** means a

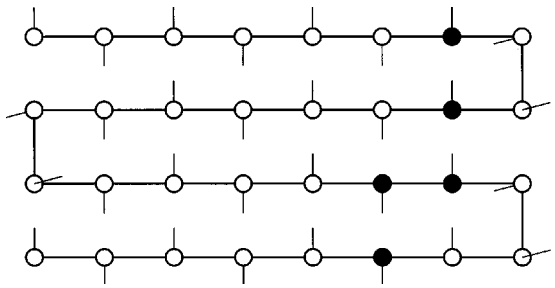

FIG. 6. Example of a native three-turn β sheet. Filled and open circles represent ***A*** and ***B***, respectively.

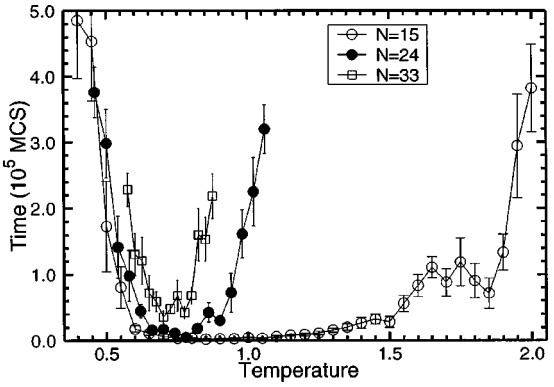

FIG. 7. Folding time τ_f versus temperature for *Sequence 1* with $N=15$, 24, and 33. Each data point is the mean for ten MC runs.

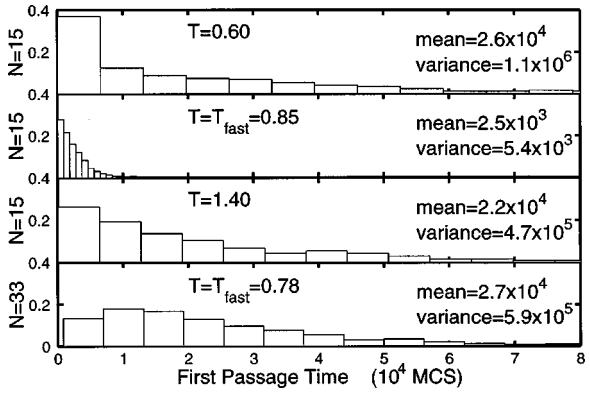

FIG. 9. Histograms showing distribution of the first passage time from unfolded initial states to the native state for chains with $N=15$ at $T=0.60$, 0.85, and 1.40 for *Sequence 1*, and $N=33$ at $T=0.78$ for *Sequence 2*. The data for each temperature were obtained by executing 1000 MC runs.

higher number of disordered states. The dependence of the fastest-folding temperature, $T_{\rm fast}$, on N is relatively weak (Fig. 8).

Analyzing the folding statistics, we have found (see, e.g., Fig. 9) that for short chains (e.g., for $N=15$) the distribution of $t_{\rm fp}$ is close to the Poissonian one at $T=T_{\rm fast}$ (because $\langle t_{\rm fp}\rangle$ is comparable with the variance about the mean) and far from Poissonian at higher and lower temperatures. For long chains (e.g., for $N=33$), the distribution of $t_{\rm fp}$ is far from Poissonian at all temperatures.

Simulating folding at different temperatures, we kept track of the time the chains spent in the native state. For example, Fig. 10 shows the mean occupation of the native state versus temperature for the chains with $N=15$, 24, and 33, for *Sequence 1*. The results are nearly independent of the chain length (for *Sequence 2*, the data are similar). In particular, the folding temperature, $T_{\rm fold}$, corresponding by definition to 50% occupation time of the native-state, is about 0.47 for all the chain lengths. The fastest-folding temperature, $T_{\rm fast}\simeq 0.80$ (Fig. 8), is in fact also almost independent of N. Thus, we have $T_{\rm fast}>T_{\rm fold}$. For comparison, we may note that for the 2D and 3D Go models[14] the dependence of $T_{\rm fast}$ on N is weak as well (actually, $T_{\rm fast}$ slightly increases with increasing N), but in contrast to our results the Go model predicts $T_{\rm fold}\simeq T_{\rm fast}$ and $T_{\rm fold}>T_{\rm fast}$ in the 2D and 3D cases, respectively.

In addition, analyzing the folding kinetics below $T_{\rm fold}$, we have calculated the glass transition temperature, T_g. By definition,[11] T_g is the temperature at which $\tau_f=(\tau_{\rm max}+\tau_{\rm min})/2$, where $\tau_{\rm max}$ is the relevant maximum time and $\tau_{\rm min}$ is the fastest folding time. In MC simulations, $\tau_{\rm max}$ may be identified with the duration of runs. In our case, $\tau_{\rm max}=10^8$ MCS (here it is appropriate to recall that in our simulations one MCS corresponds to N trials of monomer moves; thus, the real duration of MC runs, corresponding to a given value of $\tau_{\rm max}$ is proportional to N). With this (somewhat arbitrary) specification, we got $T_g=0.35$, 0.40, and 0.43 for *Sequence 1* with $N=15$, 24, and 33, respectively (the results for *Sequence 2* are similar). Comparing $T_{\rm fold}$ with T_g, we obtain $T_{\rm fold}/T_g=1.34$, 1.18, and 1.09, respectively. The ratio $T_{\rm fold}/T_g\geqslant 1$ is sometimes considered[11] to be an indicator of ''good folders.'' In this sense, all our chains are good folders.

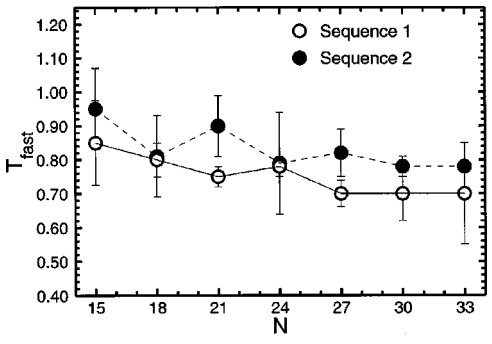

FIG. 8. Fastest-folding temperature as a function of the chain length.

The data shown in Figs. 7, 8, and 9 characterize folding globally. Additional information on the folding kinetics is obtained by scrutinizing typical folding trajectories. As an example, we show the chain energy during folding for $N=15$ (Fig. 11) and 33 (Fig. 12) at $T=0.50$ ($T\simeq T_{\rm fold}$), 0.60 ($T_{\rm fold}<T<T_{\rm fast}$), and $T=T_{\rm fast}=0.85$ and 0.78, respectively. For these two chains, the minimum energy of the native state, $E_{\rm min}$, i.e., the energy corresponding to the most favorable orientation of monomers, is equal to -16 and -43. At relatively low temperatures ($T=0.50$ and 0.60), the folding is seen to be hindered due to trapping into metastable states

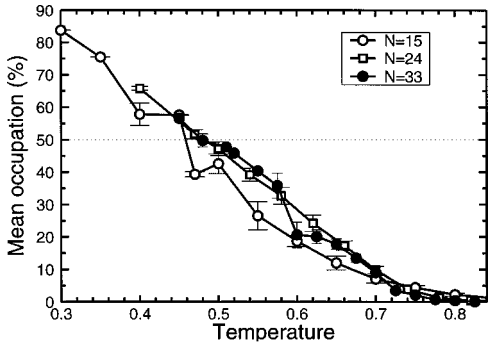

FIG. 10. Mean occupation of the native state as a function of temperature for the chains with $N=15$, 24, and 33, for *Sequence 1*. The dashed line defines the occupation corresponding to the folding temperature. (For *Sequence 2*, the results are similar.)

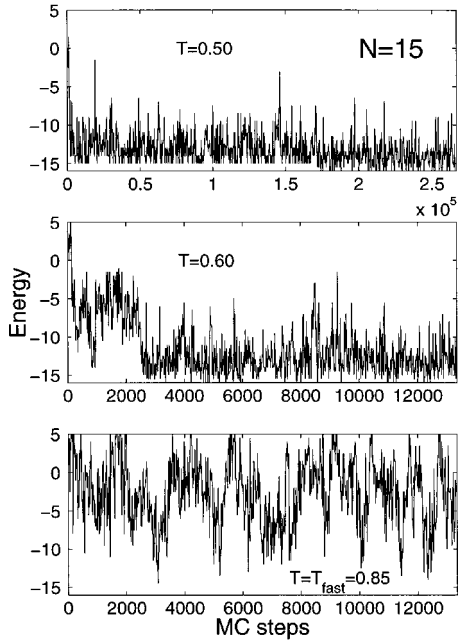

FIG. 11. Chain energy as a function of time for $N=15$ for *Sequence 1* at $T=0.50$ (top), 0.60 (middle), and 0.85 (bottom). The asymptotic occupation of the native state in these cases is ~45%, 20%, and 2%, respectively.

(see, e.g., the top panels of Figs. 11 and 12 at $t \leqslant 1.75\times10^5$ and 3.6×10^5 MCS, respectively). The time required for the chain to reach $E_{\min}$ is in this case close to the mean folding time (Fig. 7). With increasing temperature up to T_{fast}, the average chain energy and the noise in the kinetic curves become much higher (note that in this case the energy almost always exceeds $E_{\min}$) and the native structure is often

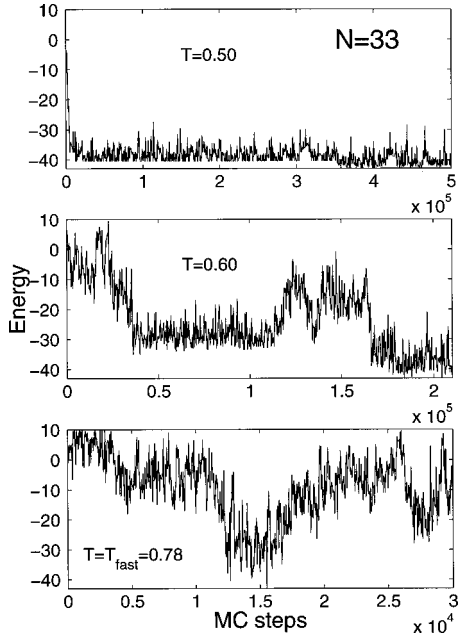

FIG. 12. Chain energy as a function of time for $N=33$ for *Sequence 2* at $T=0.50$ (top), 0.60 (middle), and 0.78 (bottom). In these cases, the asymptotic occupation of the native state is ~50%, 25%, and 3%, respectively.

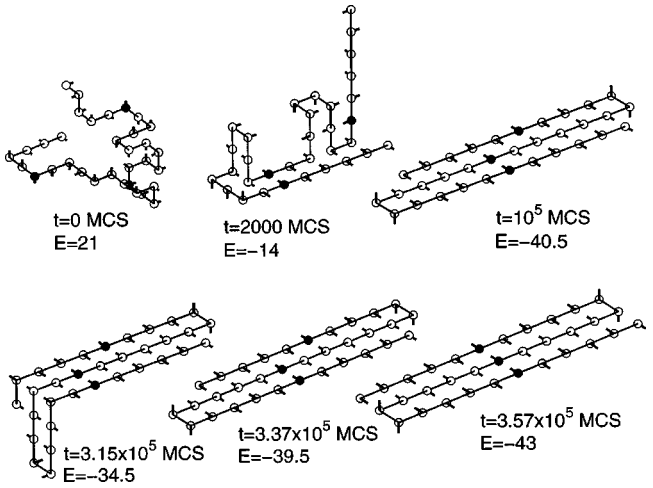

FIG. 13. Typical chain structures corresponding to the trajectory exhibited on the top panel of Fig. 12. Initially (at $t=0$), the chain is in the random coil state. At $t=2\times10^3$ MCS, the chain contains embryons of a β sheet. At $t>3\times10^4$ MCS, the chain is in the metastable state which is very similar to the native state. The latter state is reached at 3.57×10^5 MCS.

reached without formation of all the hydrogen bonds possible in this structure.

To illustrate explicitly the type of metastable states on the way to the native state, we show (Fig. 13) typical chain structures corresponding to the trajectory exhibited on the top panel of Fig. 12. The embryos of a β sheet are seen to be formed rather rapidly (already at $t=2\times10^3$ MCS). Later on (at $t>3\times10^4$ MCS), the chain is trapped to a state which is very similar to the native state (the only difference is that one of the sides is shifted one lattice spacing). In this state, the chain spends a rather long period of time before reaching the native state at $t\simeq3.6\times10^5$ MCS.

C. Folding time vs chain length

In earlier works,[12,14] where proteins were represented as a completely flexible heteropolymer with isotropic monomer–monomer interactions, the dependence of the mean folding time on the chain length was found to be described by Eq. (1) with an interaction- and temperature-dependent exponent λ. For example, at $T=T_{\text{fast}}$, λ was demonstrated[12] to be $\simeq6$, 4, and 2.7 for chains with a random choice of monomer-monomer interactions, for sequences designed to provide a pronounced energy minimum of their ground-state conformation, and for the Go model with the ''ideal'' design of sequences, respectively. With decreasing temperature, λ is known to increase.[13,14] For the simplest ideal β sheets with only one turn, the model,[13] taking into account ODI and the rigidity of the polypeptide chain, predicts $\lambda\simeq6.8$ at $T<T_{\text{fold}}$.

In our present study, the ODI interactions are similar to those used in Ref. 13. In addition, our model incorporates isotropic interactions. The native state chosen in our simulations is a β sheet with two turns (Fig. 4). For this model, we have studied in detail the dependence of τ_f on N at $T=T_{\text{fast}}$ and $T=0.47\simeq T_{\text{fold}}$ for *Sequences 1* and *2* with N from 15 to 33. For each N, we used 20 MC runs in order to get τ_f. The results obtained were fitted to Eq. (1). At T

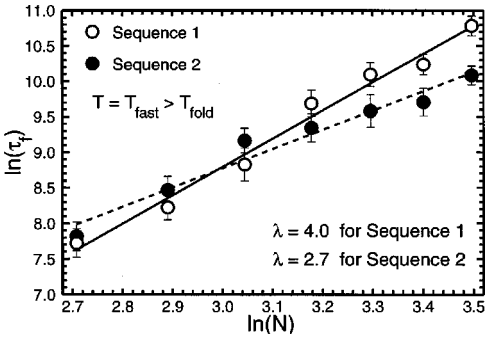

FIG. 14. Mean folding time versus chain length for *Sequences 1* and *2* with $N=15$, 18, 21, 24, 27, 30, and 33 at $T=T_{\text{fast}}>T_{\text{fold}}$ (for T_{fast}, see Fig. 8).

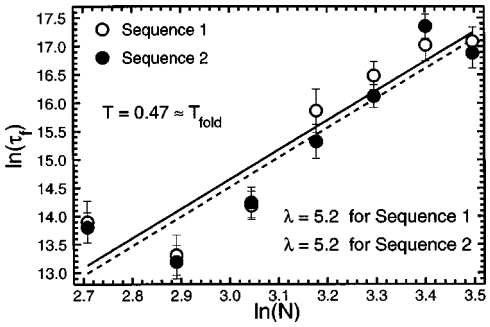

FIG. 15. Mean folding time versus chain length for *Sequences 1* and *2* with $N=15$, 18, 21, 24, 27, 30, and 33 at $T=0.47\simeq T_{\text{fold}}$ (for T_{fold}, see Fig. 10).

$=T_{\text{fast}}$, the fit is quite good and the dependence of τ_f on N for *Sequence 1* is found (Fig. 14) to be slightly stronger than that for *Sequence 2*. In these cases, $\lambda=4.0\pm0.4$ and 2.7 ± 0.4, respectively. These exponents are close to that obtained[12] at $T=T_{\text{fast}}$ for the heteropolymer sequences with a pronounced energy minimum of the native state and for the Go model with the ''ideal'' design of sequences. At $T=0.47\simeq T_{\text{fold}}$ (Fig. 15), the fit is ''poorer'' at low N, because the folding process seems to be more sensitive to the chain heterogeneity. The exponent, $\lambda=5.2\pm0.8$ (for both *Sequences*), obtained in the latter case, is somewhat lower than that[13] for the simplest one-turn β sheet at $T<T_{\text{fold}}$.

The exponent $\lambda=5.2$ found in our study at $T\simeq T_{\text{fold}}$ is rather large, and accordingly it is of interest to speculate whether this value is physically realistic. To address this point, we need to estimate τ_f predicted for real proteins by Eq. (1) with $\lambda=5.2$. Such an estimate is not straightforward, because in simulations we use MC units for time. To evaluate τ_f in real units, we rewrite Eq. (1) as

$$\tau_f=\tau_f^m(N/N_m)^{\lambda}, \tag{3}$$

where N_m is the minimum value of N where the power law makes sense, and τ_f^m is the folding time for the chain with $N=N_m$. For our model, we may use $N_m=6$, because at $N<6$ the β sheets under consideration simply cannot be formed. For $N=N_m$, the chain reconfiguration is not expected to limit the folding kinetics (because the chain is short), and accordingly the folding can be treated as a simple monomolecular process with the Arrhenius rate constant,

$$k_m\equiv 1/\tau_f^m=\nu\exp(-E_a/k_BT), \tag{4}$$

where $\nu=10^{12}-10^{13}$ s^{-1} is the ''normal'' preexponential factor, and $E_a=10-20$ kcal/mol is the activation energy. Employing $\nu=10^{13}$ s^{-1}, $E_a=10$ kcal/mol, and $T=300$ K, we obtain $\tau_f^m=2\times10^{-6}$ s. Most real proteins contain 200–600 amino-acid residues, but only part of them participates in formation of β sheets. For large β sheets, we may for a rough estimate use $N=100$. Substituting all these values into Eq. (3) yields $\tau_f=4$ s. The time obtained is thus rather long but nevertheless realistic (see, e.g., Ref. 3).

IV. CONCLUSION

We have shown that combination of the ODI and heteropolymer models opens up new possibilities in simulations of secondary protein structures. In particular, the formation of β sheets occurs if the directed (hydrogen) bonds dominate the isotropic interactions characteristic of the pure heteropolymer model. We have mimicked various kinds of antiparallel β sheets and studied the dependence of the folding time of such sheets on the chain length and temperature. With other choices of values of the model parameters, our approach can be employed to design double β sheets, or structures containing single or double β sheets and heteropolymer-type domains.

At the most interesting temperatures, the dependence of the folding time of two-turn β sheets on the chain length is found to be described by the power law (1) with $\lambda=5.2$. This value is slightly higher than that ($\lambda=4$) obtained by Gutin and co-workers[12] at $T=T_{\text{fast}}$ for the heteropolimer sequences with a pronounced energy minimum of the native state, but lower than that[13] ($\lambda=6.8$) for the simplest one-turn β sheets. We believe that our present results obtained for two-turn β sheets are more representative compared to those for one-turn β sheets, because in reality the folding of the latter sheets will be much more perturbed by other protein domains. The fact that for two-turn β sheets λ is slightly higher that that for the 3D heteropolymers seems to be connected with the dimensionality of target structures. In our case, the model is three-dimensional but the target structures are two-dimensional. In the purely 2D case, $\lambda=6-7$ even for the Go model.[14] Thus, our case is somewhere between the 3D heteropolymer model and the purely 2D models. The difference between our results and those obtained by Gutin *et al.*[12] is however not significant. Thus, the values $\lambda\simeq4-5$ are expected to be what one can obtain from the lattice models with the most realistic choice of monomer–monomer interactions for globular proteins or proteins containing β sheet domains. Somewhat lower λ are expected for proteins consisting of bundles of α helices, because in the latter case the target structures are more accessible.[22] The latter conclusion is in agreement with the experiment[23] indicating that the folding of α helices is much faster than that of β sheets.

ACKNOWLEDGMENTS

One of us (K.D.) thanks E. Reimhult for collaboration on the initial stages of this project. Financial support for this

work has been obtained from TFR (Contract 251-98-796) and from the SSF Biocompatible Materials Program (Contract A3 95:1).

[1] M. Karplus, Folding Des. **2**, S69 (1997).
[2] E. I. Shakhnovich, Folding Des. **3**, R45 (1998).
[3] D. V. Laurents and R. L. Baldwin, Biophys. J. **75**, 428 (1998).
[4] M. Gruebele, Annu. Rev. Phys. Chem. **50**, 485 (1999).
[5] N. D. Socci, J. N. Onuchic, and P. G. Wolynes, Proteins **32**, 136 (1998).
[6] H. S. Chan and K.A. Dill, J. Chem. Phys. **100**, 9238 (1994).
[7] C. J. Camacho and D. Thirumalai, Europhys. Lett. **35**, 627 (1996).
[8] A. Sali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).
[9] H. S. Chan, Nature (London) **373**, 664 (1995).
[10] D. K. Klimov and D. Thirumalai, Phys. Rev. Lett. **76**, 4070 (1996).
[11] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **101**, 1519 (1994).
[12] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, Phys. Rev. Lett. **77**, 5433 (1996).
[13] V. P. Zhdanov, Europhys. Lett. **42**, 577 (1998).
[14] M. Cieplak, T. X. Hoang, and M. S. Li, Phys. Rev. Lett. **83**, 1684 (1999).
[15] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell* (Garland, New York, 1994).
[16] C. Branden and E. J. Tooze, *Introduction to Protein Structure* (Garland, New York, 1999).
[17] V. P. Zhdanov and B. Kasemo, Proteins **29**, 508 (1997).
[18] D. Hoffmann and E.-W. Knapp, J. Phys. Chem. B **101**, 6734 (1997).
[19] D. K. Klimov and D. Thirumalai, Phys. Rev. Lett. **79**, 317 (1997).
[20] S. Takada, Z. Luthey-Schulten, and P. G. Wolynes, J. Chem. Phys. **110**, 11616 (1999).
[21] Y. Zhou and M. Karplus, Nature (London) **401**, 400 (1999).
[22] V. P. Zhdanov and B. Kasemo (unpublished).
[23] V. Munoz, P. A. Thompson, J. Hofrichter, and W. A. Eaton, Nature (London) **390**, 196 (1997).