Temperature dependence of the folding rate in a simple protein model: Search for a ''glass'' transition

A. Gutin, A. Sali,[a] V. Abkevich, M. Karplus, and E. I. Shakhnovich
Harvard University, Department of Chemistry and Chemical Biology, 12 Oxford Street, Cambridge, Massachusetts 02138

(Received 7 July 1997; accepted 16 January 1998)

Monte Carlo simulation of model proteins on a cubic lattice are used to study the thermodynamics and kinetics of protein folding over a wide range of temperatures. Both random sequences and sequences designed to have a pronounced minimum of energy are examined. There is no indication in the kinetics of a ''glass'' transition at low temperature, i.e., below the temperature of the equilibrium folding transition, the kinetics of folding is described by the Arrhenius law at all temperatures that were examined. The folding kinetics is single-exponential in the whole range of studied temperatures for random sequences. The general implications of the temperature dependence of the folding rate are discussed and related to certain properties of the energy spectrum. The results obtained in the simulations are in qualitative disagreement with the conclusions of a theoretical analysis of protein folding kinetics based on certain kinetics assumptions introduced in the Random Energy Model. The origins of the discrepancies are analyzed and a simple phenomenological theory is presented to describe the temperature dependence of the folding time for random sequences. © *1998 American Institute of Physics.* [S0021-9606(98)52215-5]

I. INTRODUCTION

The protein folding problem has both thermodynamic and kinetic aspects. The existence of a unique structure for each globular protein demonstrates that this structure is thermodynamically stable and kinetically accessible in a biologically reasonable time under physiological conditions.[1] The kinetic aspect of the problem, which is the concern of the present paper, is often phrased in terms of the Levinthal paradox; i.e., the difficulty of finding the unique native state by searching through the astronomically large number of conformations that exist for a polypeptide chain. It has been realized that the original statement by Levinthal does not provide a full description of the problem. For example, although the helix-coil transition also involves a very large number of conformations, rapid folding is a direct consequence of the fact that only local interactions are involved.[1,2] The long-range interactions, which lead to the cooperative folding transition, are an essential element in the difficulty of the search problem.[1,3–5]

The nature of the potential surface has to be considered in any analysis of the folding reaction.[1] If the potential surface were a simple well (a ''funnel'') with an energy that decreases rapidly enough to overcome the entropy reduction associated with folding, there would be no Levinthal paradox. Various suggestions for resolving the paradox have stressed the importance of local interactions. These include the introduction of specific structural entities (helixes, sheets, zippers)[6–8] as well as more general constructs, such as the ''harmony principle''[9] and the ''principle of minimum frustration''.[10] The importance of nonlocal interactions for fast folding at the conditions when the native state is stable, has been pointed out.[5,11,12]

Recent theoretical studies based on heteropolymer lattice simulations have introduced a new perspective on the protein folding problem.[1,13–15] Because of the complexity of proteins and their folding reactions such simplified models can serve to obtain insights that are not yet available from experiment. It has been demonstrated with heteropolymer models, that the thermodynamics of the protein folding transition can be studied analytically. The replica mean-field approach has been used to investigate various models of infinite heteropolymers, from the simplest case where the interaction energies between monomers were assumed random and uncorrelated[16–18] to more realistic model treatments where heteropolymers were characterized by the monomer sequence.[19–21] It was concluded[16,19,21] that the thermodynamic behavior of such heteropolymers can be adequately described by the Random Energy Model (REM) of Derrida.[22] This led to important analytic results, some of which have been applied to the thermodynamics of protein folding.

Garel and Orland[17] called these results into question when they concluded that the REM is a good approximation for random heteropolymers only in high-dimensional space ($d \geqslant 4$), while in three dimensions the thermodynamics of heteropolymers cannot be described by the REM. Their analysis was based on a simplified model where the polymer bonds were not treated explicitly; i.e., monomers were positioned on a simplex.[17] The analysis of microscopic models, where polymer bonds were taken into account explicitly,[16,19,21] indicates that the REM is a good approximation for the thermodynamics of three-dimensional heteropolymers. Further, analytic results for this model have

[a]Present address: The Rockefeller University 1230 York Avenue New York, NY 10021-6399.

been confirmed by lattice simulations of relatively short heteropolymer chains (e.g., a 27-mer on a three-dimensional cubic lattice[23] and a 16-mer on a two-dimensional square lattice[24]).

Of particular interest for the protein folding problem is the conclusion from the random energy model that there exists a critical temperature (T_c) below which a sequence with sufficient heterogeneity has a stable, essentially unique structure. This satisfies the thermodynamic requirement for protein folding. Lattice simulations have suggested that a significant fraction of random sequences have such a stable unique structure below T_c.[25] Further, full enumeration of the 16-mer conformations on a two-dimensional square lattice has demonstrated that the phase diagram as a function of temperature and the average attraction between monomers consists of an extended coil region, a homopolymer-like disorganized globule region and an organized frozen globule, which corresponds to the lowest energy (native) conformation and is stable below T_c. The phase diagram from exact enumeration agrees well with analytical heteropolymer theory.[16]

These and other results from heteropolymer theory for protein thermodynamics suggest that a corresponding model should be useful for studying the kinetics of protein folding. However, it has not been possible to obtain analytic results for the kinetics. Bryngelson and Wolynes (BW)[26] introduced assumptions not inherent in the random energy model to permit an approximate analytic treatment of folding kinetics. They assumed that the kinetics can be deduced by the use of a Metropolis Monte-Carlo (MC) algorithm in ''energy space.'' Since the phenomenological random energy model does not have geometric features, the Monte-Carlo moves could not be based on geometric properties of the polymer chain, i.e., there is no connection in the random energy model between geometric properties of a conformation and its energy. Consequently, an additional assumption concerning the possible energy changes associated with a move was required. BW assumed that at each MC step the energy of the ''attempted'' conformation is statistically independent of the energy of the existing conformation. This means that the kinetic scheme used by them in extending the random energy model permitted any change of energy in an attempted move, with equal probability. For standard Monte-Carlo simulations in real space, small energy changes are most likely at each step because the ''attempted'' and ''existing'' structures are geometrically similar.

An important conclusion from the theory of BW[26] was that two temperatures play a key role in determining the folding properties of a heteropolymer sequence. One is the critical temperature, T_c, already discussed. They denote T_c by T_g because they suggest that it corresponds to the glass transition temperature for the heteropolymer; we use T_c in what follows. The BW model leads to the conclusion that below T_c the heteropolymer is frozen into one of many random low-energy conformation because it does not have enough energy to overcome the barrier separating such conformations. The other significant temperature introduced in BW is the folding temperature, T_f, which corresponds to the midpoint of the thermodynamic transition between the native and denatured state; it is the same as T_m in the protein literature. BW concluded that T_f must be higher than T_c for folding to occur on a reasonable time scale; below T_c, the time required for folding was stated to be equal to Levinthal ''time'' (i.e., the effectively infinite time required for an unbiased random search of all significant conformations).

Since the thermodynamic criterion for a stable unique ground state in the random energy model requires $T_f < T_c$, as indicated above, random sequences cannot fold in the BW model; i.e., the temperature, T_f at which such random sequences can fold ($T_f > T_c$) would lead to an unstable ground state.[27] To overcome this difficulty, BW proposed that protein sequences have specific biases toward the native state that make folding possible. The existence of such biases on the entire potential energy surface, which would result in an energy ''funnel'' leading toward the native state,[28] is referred to as ''the principle of minimum frustration,'' which is closely related to the consistency or harmony principle of Go *et al.*[29] Other kinetics treatments based on the REM have been proposed.[30–33] The kinetic assumptions in Refs. 31–33 were similar in spirit to the ones used by BW[26] and the results of Refs. 32,33 were qualitatively in accord with the ones obtained by BW.

Lattice Monte Carlo simulations of a 27-mer on a three-dimensional cubic lattice with random interactions[34,35] showed that only a subset of random sequences fold in a reasonable time in the neighborhood of the transition temperature T_f; out of a total of 200 sequences whose folding was studied, only 30 folded rapidly as defined by cutoff in the number of Monte-Carlo steps. The only difference between random sequences and the subset of total folding sequences was that the latter satisfied the thermodynamic requirement for protein folding at a higher temperature due to the presence of a large energy gap between the lowest energy (native) state and the low-energy excited states (not similar structurally to the ground state). This led to a folding temperature (T_f) above the critical temperature (T_c) for these selected sequences, in agreement with the conclusion of BW. An analysis of the kinetics of folding for these sequences showed that the distribution of folding times was exponential over a wide temperature range, but that the temperature dependence was not Arrhenius-like; instead a bell-shaped curve for the logarithm of the rate versus the inverse temperature was obtained; i.e., although the rate increased with temperature at low temperature, it decreased again at higher temperatures[36,3] Because of strong ($2kT$) average attraction between monomers, which had been introduced to ensure that the native state is fully compact (i.e., belongs to the enumerated set of conformations in a $3\times3\times3$ cube), the folding rate was too slow to examine what happens significantly below T_c.[35,36]

Other lattice simulations have shown a corresponding temperature dependence for the folding rate.[37,3] Socci and Onuchic[37] made a study of folding kinetics of the 27-mer at different temperatures and looked for a glass transition. They introduced an ''operational'' glass transition temperature as the temperature at which the folding ''time'' exceeds 10^9 Monte-Carlo steps and found slow folding that satisfied this criterion at low temperatures. However, they did not demon-

strate that folding was significantly slower than that expected from the Arrhenius equation at any temperature. In the study of folding kinetics of lattice model proteins Pande *et al.*[38] observed Arrhenius dependence of folding rate on temperature. Chan and Dill[39] numerically solved the master equation describing transitions among all the conformations of a short chain on a square lattice. This allowed them to obtain the folding time even for very low temperatures, and perfect Arrhenius-like behavior was still observed. There was no indication of a kinetic glass transition. An Arrhenius-like behavior was also observed for off-lattice folding of a heteropolymer model.[40] The question of whether the low-temperature kinetics of 27-mer lattice model can be fitted by Arrhenius or non-Arrhenius law was also discussed in a recent publication.[41] The authors of Ref. 41 noted that the ''dynamic ruggedness'' of the energy landscape is much less than can be expected from the kinetic REM model. However, the temperature range at which kinetics were studied in Ref. 41 was not sufficiently broad to allow a detailed analysis of the low-temperature behavior and its implications for the concept of kinetic glass transition in protein models. The kinetic glass transition was defined in Ref. 41 as temperature at which folding is significantly slower than fastest observed rate. Such definition of kinetic glass transition makes it difficult to search for specific features that distinguish ''glassy'' behavior from obvious slowing down of folding at low temperature which can be predicted from Arrhenius law and which is a feature of any dynamics that involve energetic barrier crossing.

To explore further the question of the existence of a kinetic glass transition in heteropolymer lattice models, it is necessary to have faster folding sequences that can be studied significantly below T_c. In the present paper, this is obtained by elimination of the strong overall attraction between the monomers. As a result the native states for random sequences are not fully compact. To make sure of the generality of the results we also studied one sequence for which a moderate average attraction ($\approx kT$) between monomers was introduced, in order to obtain a maximally compact native conformation. Both sequences selected at random and designed sequences are examined. Rather than using random pairwise interactions as in the original 27-mer studies,[42,34,35] the Miyazawa-Jernigan amino acid parameters are employed[43] in specifying the sequences and the interactions between residue pairs. Analysis of the thermodynamic behavior of these sequences allows us to determine T_c. By doing simulations below T_c, we are able to make a direct test of the theoretical predictions concerning ''glassy'' folding dynamics below T_c.

It is found that even at very low temperatures ($T<T_c$), the folding rate obeys the Arrhenius equation. As pointed out by Angell in his comprehensive review of glass forming liquids,[44] ''The almost universal departure from the familiar Arrhenius law is perhaps the most important canonical feature of glass forming liquids.'' In addition he pointed to non-exponential relaxation as another attribute of a glass. Since neither of those are evident in the presented folding simulation, there is clearly no evidence for a kinetic glass transition at or below T_c. Further, the folding time at or below T_c is much shorter than the Levinthal time. This suggests that a modifications of the BW model based on REM is required to apply it to the kinetics of heteropolymer and protein folding. In this paper a simple phenomenological model is proposed to describe the temperature dependence of the folding kinetics of random sequences. The simulation results and the model are used as a basis for a general analysis of the temperature dependence of the rate of protein folding.

II. THE MODEL

The model which we used was described in detail in previous publications.[42,45,3] A protein is modeled by a self-avoiding heteropolymer chain on a cubic lattice. A monomer corresponds to an amino acid residue of a protein, as in the Miyazawa-Jernigan model. A monomer can occupy any site of the lattice; two or more monomers cannot occupy the same site. Monomers connected by a bond occupy nearest neighbor sites.

The energy of a conformation is given by

$$E=\sum_{1\leqslant i<j\leqslant N} U(\xi_i,\xi_j)\Delta_{ij}, \tag{1}$$

where the sum is taken over all pairs of monomers; $\Delta_{ij}=1$ if monomers i and j are in contact with each other, and $\Delta_{ij}=0$ otherwise. The energy $U(\xi_i,\xi_j)$ of a pairwise contact depends on the identities of monomers i and j. The values of $U(\xi_i,\xi_j)$ are taken from Table VI of Ref. 43.

The motion of a chain is simulated by the standard Monte-Carlo technique with the move set including corner flips and crankshaft moves by 90 and 180 degrees.[46] At each step a monomer is picked randomly and its possible moves (corner flip or crankshaft in a random direction) are attempted. The directions of crankshaft moves are chosen randomly with equal probability.

III. NUMERICAL RESULTS

A. Nondesigned sequences

We generated ten random amino acid sequences of 27 residues. All sequences have the same composition, which was chosen arbitrarily. The sequences are listed in Table I.

The relative energy of the native conformation $E_{\rm rel}$ is defined as[47,48]

$$E_{\rm rel}=\frac{E_{\rm nat}-E_{\rm av}}{\sigma}, \tag{2}$$

where $E_{\rm av}$ is the average energy of non-native conformations. To estimate this value, we first compute the energies of all topologically possible contacts between all monomer pairs. From this we calculate the average energy $e_{\rm av}$ of a contact, and then estimate the average energy of non-native conformation as $E_{\rm av}=C\cdot e_{\rm av}$, where C is the number of contacts in the native conformation. In Table I we give the normalized value $e_{\rm rel}=E_{\rm rel}/C$. This value serves as a measure of the ''energy gap.''

TABLE I. Random sequences of 27 monomers. Energy of the native conformation E_{nat}, the relative energy of the native conformation E_{rel}/C (normalized by the number of contacts), the number of contacts C in the native conformation, and the MFPT at $T=0.16$ in Monte-Carlo steps are given for each sequence, except sequence #5.

#	Sequence	E_{nat}	E_{rel}/C	C	MFPT
1	DCSATYNFVPAGLSQHMRTEIEGWVKL	−5.92	−1.15	22	$5.4\cdot 10^5$
2	ENHKGLTVDAPIASYWLQTEVRGMFCS	−6.50	−1.12	22	$3.9\cdot 10^5$
3	PALETMDSFQWRCISVYGAHVLGNTKE	−6.45	−1.11	21	$3.7\cdot 10^5$
4	VKAMRLAVPLFESESNYCWGHIQTDTG	−6.32	−1.34	18	$8.6\cdot 10^5$
5 [a]	EVPSLNMHESQAFGYLRTDCGTIKVWA	−13.67	−0.73	28	$6.6\cdot 10^6$
6	PGALKDIFNYVQSGRECTEHVTMWASL	−5.61	−1.18	20	$1.8\cdot 10^6$
7 [b]	LQIVADTSNHGERMVTCAPWFSKELGY	−5.80	−1.12	20	$1.1\cdot 10^8$
8	GSRPGAFNIVMQKCDTVLWEYASTHLE	−5.68	−1.09	21	$3.0\cdot 10^5$
9	NKECIYLDPWHTGQRSTFALVGASVEM	−5.55	−1.13	21	$6.5\cdot 10^6$
10	LYSLTGTKSWQGAEEVMHCADRFINVP	−7.22	−1.30	20	$4.8\cdot 10^5$

[a]Nonspecific attraction potential of −0.3 is introduced to make the native state of this sequence maximally compact. The folding rate for this sequence is quoted at $T=0.25$.
[b]This sequence has double degenerate native state hence its slower folding (see the text).

For each of the sequences one Monte-Carlo simulation of 10^9 steps starting from random coil conformations was performed and the conformation with the lowest energy was determined. This is identified as the native structure for that sequence. Figure 1 shows the native conformations for some of these sequences. Native states for sequences #1, #4, #8 shown in Fig. 1 have fewer contacts in the native state than the 28 corresponding to a fully compact cube, while sequence #5 has fully compact native state shown in Fig. 1(d). To test that the putative native conformation is, in fact, the state of lowest energy, 25 runs were performed for each chain until it first reaches that conformation. During these runs, no conformations with lower energy were observed. This test suggests that the chosen native conformation for each sequence has the lowest energy, at least among all kinetically accessible conformations. Relatively fast folding (in less than 10^9 steps) of all the random sequences to their respective ''native'' conformations can be explained by the absence of strong average attraction between monomers which makes the motion of the chain less constrained, compared to previous studies where fast collapse to a quite dense conformation preceded folding.[35]

From 25 Monte-Carlo runs at a temperature $T=0.16$, where T is in the same units as the MJ parameters ($k_B=1$ and is dimensionless), the mean first passage time (MFPT) for reaching the native conformation was estimated for the ten sequences. The results are given in Table I. The MFPT values for this temperature vary by a factor of about 30. Sequence #7 folded significantly more slowly than the other sequences. The reason for the slow folding of sequence #7 turned out to be an almost exact double degeneracy of the native state: the lowest energy conformation has energy of

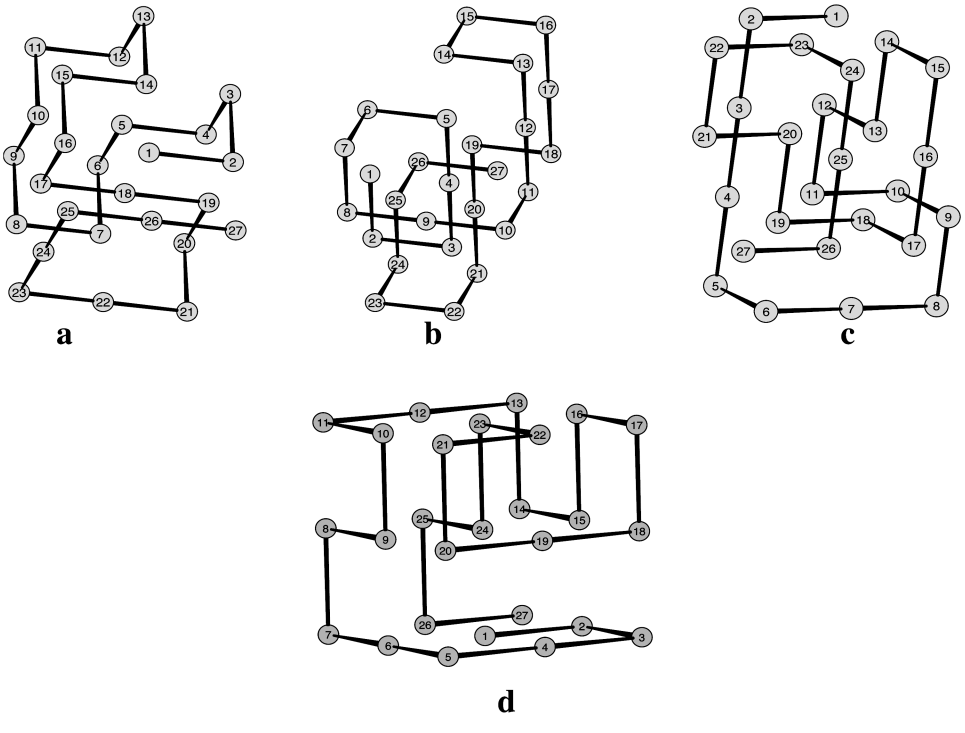

FIG. 1. Native conformations that is conformations with the lowest energy for (a) sequence #1, (b) sequence #4, and (c) sequence #8.

-5.80 while there exists also a second conformation with E$=-5.75$ (i.e., only 0.3 kT higher than $E_{\rm nat}$ for that sequence). At $T=0.16$ sequence #7 folds fast, with MFPT of 6.7×10^6 steps to the conformation with energy -5.75 and then slowly (in 10^8 steps) interconverts to the native conformation. This phenomenon of ''kinetic partitioning'' is interesting as a possible simple model of prion behavior. The temperature dependence of the folding rates into two ground states of the ''prion-like'' sequence was studied in Ref. 49 and was found to be qualitatively similar to the temperature dependence of folding rates of the ''normal'' sequences studied in this paper.

The prion-like behavior is atypical: Only one out of ten random sequences generated for the present study exhibits such a behavior. Since our goal here is to study temperature dependence of folding rates in typical random sequences, we chose the following sequences for more detailed analysis: sequences #1 and #4, which have intermediate folding times representative of the whole sample of random sequences, and, for comparison, sequence #8, which has the shortest folding time. To check whether our conclusions are sensitive to the degree of compactness of the native conformation, a fully compact sequence was studied; this was generated from sequence #5 by adding a nonspecific attraction of -0.3 to every pairwise interaction [see Fig. 1(d)].

We find that all the qualitative features of the folding behavior of the sequence with the fastest folding rate as a function of temperature are the same as those of the other sequences, including the one with a maximally compact ground state. This suggests that for longer polymers most randomly generated sequences will exhibit similar folding behavior. Consequently, we chose a fast-folding 48-mer for detailed study (see below) so as to be able to obtain satisfactory statistics for this system in a reasonable time. It is important to emphasize that the present calculations are time-consuming which means that a limited number of sequences can be studied in detail.

A study of the kinetic and thermodynamic behavior was performed for sequences #1, #4, #8 and modified sequence #5. (Details are given in the figure captions). Figure 2(a) shows the inverse temperature dependence of the equilibrium energy E obtained from long Monte Carlo simulations. According to the REM[22] (also see Appendix A), the average energy decreases with temperature when $T>T_c$ (the average energy varies approximately linearly with $1/T$) and becomes constant close to E_c at temperatures below T_c, corresponding to the dominance of a few conformations with energies E_c close to that ($E_{\rm nat}$) of the native state. The three sequences exhibit similar behavior and $1/T_c$ is somewhere between 8 and 10. The fact that the three sequences have approximately similar values of T_c is expected because T_c is a self-averaging quantity in the sense that its value should not depend on a particular realization of a random sequence.[22] This fact was confirmed in a lattice model study where conformations of a small chain were exhaustively enumerated.[24] The probability of a large ($\sim NT_c$) energy gap in a random sequence is very low for $N\gg1$ (where N is the number of monomers) (see Ref. 34). Therefore, for random sequences the energy of the native state $E_{\rm nat}$ is close to E_c, the energy

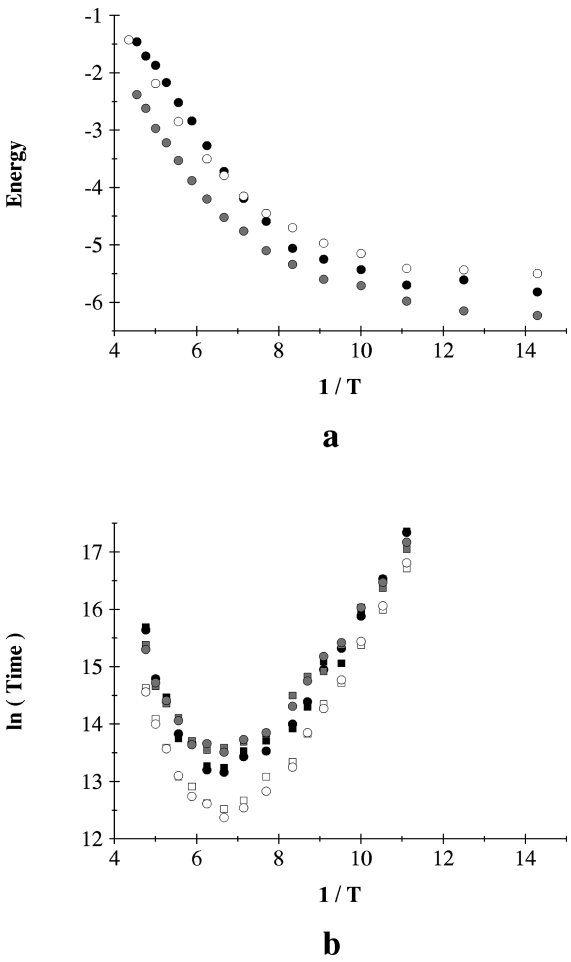

FIG. 2. Inverse temperature dependences (a) of the equilibrium energy E and (b) of the MFPT (circles) and the median time divided by ln 2 (squares) averaged over 100 runs for sequences #1 (in black), #4 (in gray), and #8 (in white).

at which the quasicontinuous part of the energy spectrum ends.[25] This makes it possible to evaluate E_c for each sequence *as the energy at which the dependence of E on T levels off*. (It should be also noted that the exact values of E_c are not critical for the analysis.)

The inverse temperature dependence of the logarithm of the MFPT for sequences #1, #4, #8 is shown in Fig. 2(b). For all three sequences there is an optimal temperature at which the rate is maximal. In the vicinity of the optimal temperature the dependence of the rate is parabolic. As was mentioned in the Introduction, such a nonmonotonic dependence of the MFPT on temperature was found in a number of previous studies.[50,37,3,36] For the present analysis, it is the low temperature behavior which is of primary interest. At low temperatures the dependence of ln MFPT on $1/T$ becomes linear which is characteristic of Arrhenius behavior. By comparing Figs. 2(a) and 2(b) we see that the temperature at which the Arrhenius dependence appears is close to the temperature of the equilibrium transition T_c for random sequences. This implies that below T_c the activation energy and entropy, as well as the equilibrium energy, do not depend on temperature, i.e., the dependence of ln MFPT on $1/T$ is well approximated by straight line at $T<T_c$, which, ac-

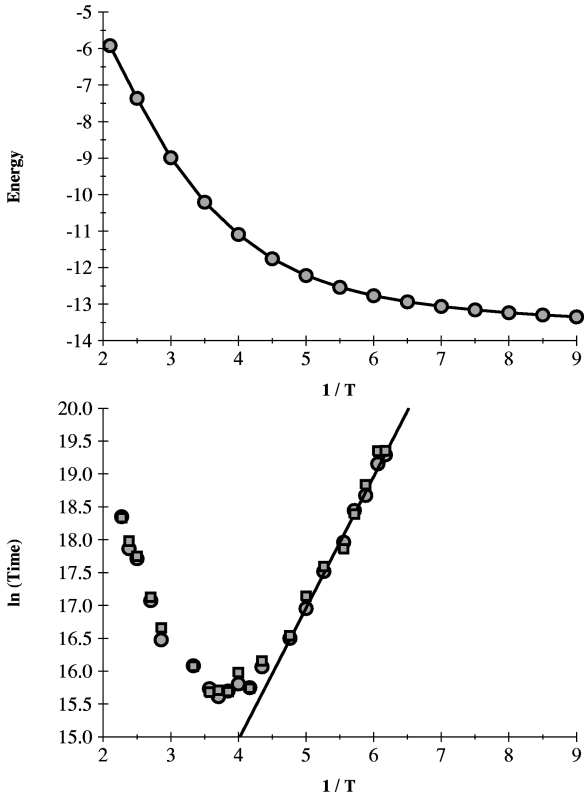

FIG. 3. Inverse temperature dependences (a) of the equilibrium energy E and (b) of the MFPT (circles) and the median time divided by ln 2 (squares) averaged over 100 runs for sequence #5. The line in (a) represents the result of the histogram method calculation, while circles are actual datapoints. The consistency attests that equilibrium has been achieved in calculation. The line in (b) is drawn to guide the eye.

cording to classical Arrhenius theory is a signature of temperature independence of activation energy and entropy.

Figures 3(a) and 3(b) show the thermodynamics and kinetics for the sequence #5 which has compact native conformation. It clearly has temperature dependent behavior that is similar to that of the noncompact sequences. There is a parabolic temperature dependence of MFPT at $T>T_c$ and almost perfect Arrhenius dependence at $T<T_c$ (in this case $T_c \approx 0.22$). Similar thermodynamic and kinetic properties as shown in Figs. 2 and 3 were obtained for the model used in Refs. 34,35, in which the native states were almost always compact and the interactions between residues were chosen to have independent random values (A.S., E.S. and M.K., unpublished results).

Another qualitative feature of ''glassy dynamics''[44] is that the distribution of folding times below T_c is nonexponential with an enhanced probability of very slow folding events due to chain trapping in deep non-native states, i.e., a signature of the kinetic glass transition is a broad spectrum of relaxation times due to the differences between molecules that are frozen in local minima below T_c.[27,44] Therefore it is important to determine whether or not the folding kinetics (distribution of first passage folding times) are single-exponential. To examine this point we show in Fig. 2(b) the inverse temperature dependence of the median folding time, the time within which a chain finds its native conformation in half of the runs. If folding is a single-exponential process, the median time is equal to MFPT multiplied by ln 2. If the distribution of folding times is a sum of exponentials, the median time is less than MFPT $\times$ ln 2. This simple method is useful because it compares two averaged quantities, each of which can be reliably determined from a modest number of runs; at each temperature 100 runs were performed to estimate the average time and median time. It is seen from Fig. 2(b) that folding of the three random sequences at all temperatures studied is not distinguishable from a single-exponential by this criterion. The same conclusion, for a system with random interresidue interactions was made in Ref. 36. based on a plot of ln k vs time. It is also clear from Fig. 3(b) that the kinetics are still single-exponential in the case when the native state is fully compact, as it is for sequence #5.

The most important aspect to be considered comparing our results with theory of BW (and its subsequent ramifications[51]), is whether or not there is any *qualitative* signature of a kinetic glass transition at T_c. It is clear from Figs. 2(b), and 3(b) that the dependence of folding time does not exhibit any peculiar features (sigmoidal or a plateau or a cusp) which may be expected in a finite system as a manifestation of a transition. Instead, the dependence of MFPT on temperature is smooth, and Arrhenius behavior is observed up to the lowest temperatures examined, considerably below T_c. Further, we showed that for all studied random sequences the kinetics are single-exponential at all temperatures, which is not consistent with kinetic glass transition.[44]

As regards the comparison of quantitative aspects of the BW theory with the simulation results, we should note that their conclusion that folding time equals Levinthal time at $T=T_c$ and stays unchanged at lower temperatures, is a consequence of averaging of folding rate over the ensemble of sequences. This way of determining folding rate is flawed because at low temperature (below T_c) the major contribution to the average (over ensemble of sequences) rate is provided by a very small number of extremely improbable sequences having exceptionally fast (barrier free) folding (a discussion of this issue is given in the Discussion section and more details are provided in Appendix B). In fact, almost no sequences in the BW model will fold in ''Levinthal time'' at T_c. Rather, a vanishingly small fraction will fold extremely fast and the vast majority will fold much more slowly than ''Levinthal time'' in the BW model. Averaging the rate over the ensemble of sequences in that model yields ''Levinthal time'' (see below). In contrast here we study folding of typical random sequences. It is not appropriate to compare our results which pertain to several typical sequences with the prediction of BW which applies to the ensemble and is not characteristic of any typical random sequence.

However, it is still instructive to compare the calculated folding times with ''Levinthal'' time. To estimate the Levinthal time for the present model, a naive approach would be to assume that the Levinthal time is the time needed to search the total number of conformations, in accord with original description of the Levinthal paradox.[52] However, this is likely to be an overestimate because many of the states may be inaccessible; e.g., the chain may undergo a partial nonspecific collapse since completely open

conformations are thermodynamically unfavorable,[53,54,35,55] so that the number of conformations is effectively reduced to the semi-compact states. Since it has been shown that the statistics of the conformational energies in such an ensemble of semi-compact heteropolymer conformations follow the REM,[16] the number of thermodynamically relevant conformations Ω can be estimated using the REM. The quantity Ω is related to the experimentally measurable ''freezing'' temperature T_c and freezing energy E_c[22] (see Fig. 2) by

$$\Omega = \exp\left[-\frac{E_c}{2T_c} \right], \tag{3}$$

i.e., the number of relevant conformations is between the number that are maximally compact and the number of all possible conformations. (See Appendix A for the derivation of this result and further discussion.)

The number of relevant conformations Ω given by Eq. (3) is much smaller than the number of all possible conformations Γ. The latter can be estimated as $\Gamma \approx \gamma^{(N-1)}$, where $\gamma \approx 4.68$[56] and N is the number of monomers in the chain; for $N=27$ we obtain $\Gamma \sim 10^{17}$. According to the REM the temperature dependence of the energy reaches plateau at the limiting value of energy E_c at $T=T_c$. Therefore, we estimate E_c as the energy at which the dependence of $E(T)$ in Figs. 2(a) and 3(a) *reaches the plateau.* Substituting $E_c \approx -6$ and $1/T_c \approx 8.5$ [see Fig. 2(a)] into Eq. (3) we obtain $\Omega \approx 10^{11}$. This is close to the estimate 10^{10} given in Ref. 35 but somewhat greater since in the present case the mean attraction of monomers is less than that in the earlier calculation. The estimate can also be compared with the MFPT at T_c from our simulations, which is approximately 10^6. We see that folding rate of random sequences with no average attraction between monomers at T_c is much faster than the Levinthal time estimated for that model.

Introduction of overall attraction makes the number of thermodynamically relevant conformations smaller (i.e., it restricts the conformational ensemble to more or less compact conformations). This factor decreases the apparent ''Levinthal time.'' In fact, the same estimate for sequence #5 (for which the interaction potential includes an average nonspecific attraction as described above) gives $\Omega \sim 10^6$. This is less than MC folding time ($\approx 6.6 \times 10^6$) even at the conditions of fastest folding. [In making the estimate for E_c from Fig. 3(a) for sequence #5 one has to subtract total nonspecific attraction energy ($-8.4 = -0.3 * 28$) from the low-temperature plateau value in Fig. 3(a), since the E_c is estimated as the difference between average energy and the energy obtained at low temperature.] That the Levinthal time estimated this way can be less than the actual folding time is due to the fact that stronger compaction constrains the chain moves (i.e., very few MC trials are accepted) and slows down motion toward the native state. This simple consideration shows that comparison of the folding time with the ''Levinthal time'' for any model is somewhat arbitrary. Thus the comparison of the folding time with the ''Levinthal time'' may be not very instructive since it depends on the details of a model and the definition of the ''Levinthal time'' (e.g., does one consider all conformations, or only compact ones, etc.) and therefore does not reflect the essential physics of the problem.

B. Designed sequences

It has been argued that protein sequences are not random; i.e., that their the native state stability is higher than that of random sequences.[10,57,35,58,42,59,60] There has been also a considerable interest in statistical analysis of real sequences to assess whether they are random or not.[61–63] The question of how to detect sequence nonrandomness is a delicate one. The first simple tests did not reveal statistically significant deviations of certain sequence characteristics, such as hydrophobicity pattern, from a random distribution.[61,62] However, recent more refined analysis indicates that the distribution of different amino acids in protein sequences are not random.[63]

It is important, therefore, to study the kinetics of folding of nonrandom sequences at different temperatures, in addition to the random sequences considered above. We generated optimized sequences with a design algorithm similar to that described in Ref. 58. In the present work we did not constrain the amino acid composition and minimized the relative value of the energy of the native conformation E_{rel} [see Eq. (2)] rather than the native state energy itself, as was done in Ref. 58. The choice of E_{rel} [Eq. (2)] as a parameter to be optimized is motivated by computational convenience since at each step E_{av} and σ can be easily evaluated without running, after each mutation, a computationally expensive search in space of denatured states. Sequences were designed with low relative energy for the three native conformations that were used in the random sequence analysis (Fig. 1).

The original MJ parameters were shifted and scaled in such a way to yield values for E_{av} and σ that were the same as for corresponding random sequences. This is important for comparison with random sequences and BW theory and also it provides a direct way of comparing the behavior of designed and random sequences having the same native conformations.

To demonstrate the effectiveness of the design procedure, we calculated the number of conformation $\nu(E,Q)$ with a given energy E and a given number of native contacts Q; the method for doing this is described in Refs. 35,48. Comparison of E_{rel} for the random and designed sequences in Tables I and II show a pronounced decrease of E_{rel} for the designed sequences.

The detailed mapping of configurational space obtained in the MC simulations allows us to assess the efficiency and the role of design. Comparison of the plots $\nu(E,Q)$ for random and designed sequences (Fig. 4) indicates indeed that the design is effective. It is clear that low energy conformations with many non-native contacts are present in the native sequences and are absent in the designed sequences. Moreover, we estimate the stability gap for the two sequences. By stability gap we mean the energy difference between the native state and lowest energy conformations belonging to the denatured state. To identify denatured states on the diagrams in Fig. 4 we assume that they have approximately 5 native contacts, i.e., denatured states have a degree of similarity to the native state corresponding to the similarity of two random conformations. We use as an estimate for E_{unfolded} the

TABLE II. Designed sequences of 27 monomers. Sequences #11, #14, and #18 are designed to minimize E_{rel} of the conformations (a), (b), and (c) in Fig. 1, respectively. The energy of the native conformation E_{nat} and its relative value E_{rel}, normalized by the number of native contacts C, are given for each sequence.

#	Sequence	E_{nat}	E_{rel}/C
11	MEYYWKGLEMAYAPWWIFKGTGILAWK	−10.11	−1.80
14	IKEMKAALWGWEMTMWKMWKTSYGETY	−10.14	−2.01
18	WWWTATKLKLKMQWEKTEGPAWMKQGT	−10.32	−1.87

lowest energy of conformations having 5 native contacts. For structure (a) in Fig. 1, the stability gap, $\Delta E = E_{nat} - E_{unfolded}$ is approximately 0.5 for random sequences and 4 for designed sequence #11. This can be seen clearly in Fig. 4. The difference is of particular significance because the standard variance of interactions, σ, is the same for the random and designed sequences.

It is worth emphasizing the importance of the definition of stability gap as the energy difference between the native state and unfolded, *structurally significantly different* conformations. Care has to be used to avoid incorrect definitions of the energy gap.[64] The use of the difference between the native state and nearest to it in energy is valid only in the realm of *fully compact* states (e.g., $3\times3\times3$ cubes[34]) where native state and the first excited state have a high probability to differ significantly in structure. This is obviously incorrect for non-compact conformations where the native and ''first excited'' state can differ by only small structural rearrangement (displacement by only one bead). The latter definition was used in the recent paper of Klimov and Thirumalai[64] and resulted in a misrepresentation of earlier results[34] which dealt specifically with ''energy gap'' defined in a fully compact ensemble of conformations. In a later paper by the same authors[65] the qualitative correlation between folding rate and energy gap in *fully compact ensemble* was shown (Fig. 22 of Ref. 65).

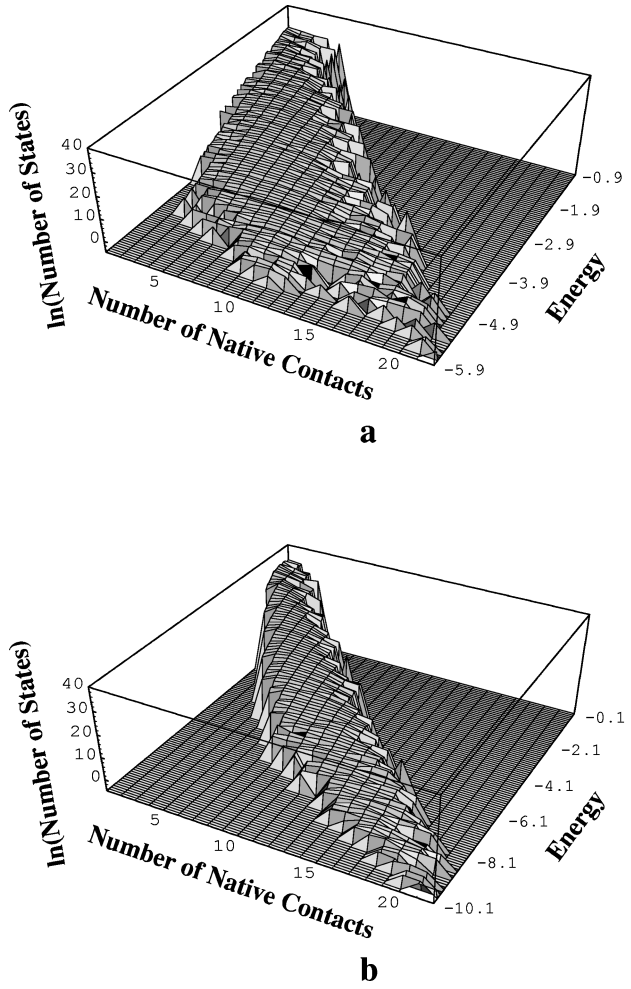

FIG. 4. Number of conformations with a given number of the native contacts and a given energy (a) for a random sequence (#1) and (b) for a designed sequence (#11).

Figure 4 shows that the random and designed sequences are similar as far as energy gap between the native state and the first ''excited state'' which differs by one monomer flip, is concerned. In other words, the energy distribution for structures in the immediate vicinity of the native state is similar for the two noncompact sequences, which have pronounced differences in their folding.

According to theory,[57,58,5] the equilibrium transition between the native and the unfolded states for designed sequences is first order, as explained in Fig. 1 of Ref. 58). In the thermodynamic limit, a first order transition corresponds to a sudden change in energy as a function of temperature. For a finite system this temperature dependence becomes sigmoidal. Such behavior is observed [Fig. 5(a)] for the designed sequences which have an inverse transition temperature $1/T_f$ at about 3.5. This should be contrasted with the behavior of a typical random sequence for which the dependence E vs $1/T$ is much less sigmoidal (see Fig. 2). Sequences selected for fast folding from the pool of 200 random sequences also showed more sigmoidal temperature folding transition than the ones that did not fold fast.[34] (The different parameter, which is the number of thermodynamically stable conformations was studied as a function of temperature in Ref. 34. It was possible to do for the model studied in Ref. 34 because the thermodynamic quantities were evaluated there for fully enumerated set of maximally compact conformations. In the present study this is not possible since no average attraction is introduced and non-compact conformations contribute significantly. To this end we use a simpler parameter, average energy E to characterize transitions in random and designed sequences.)

These sequences were more thermostable than non-folding ones, i.e., the selection for fast folding used in Ref. 34 represents a simple, ''design'' procedure and leads to the sigmoidal transition for such sequences.

The inverse temperature dependence of the MFPT for the designed sequences is shown on Fig. 5(b). The tempera-

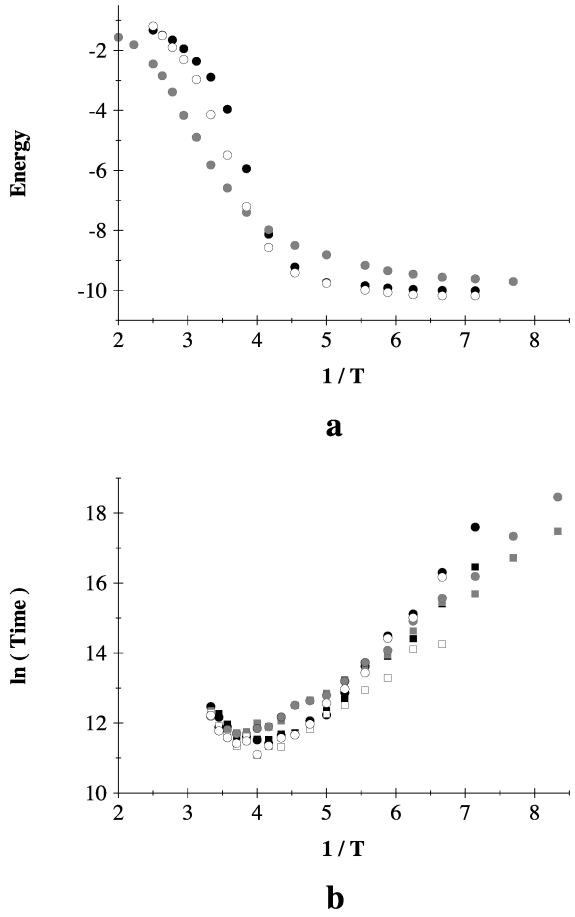

FIG. 5. The same as Fig. 2 but for sequences #11, #14, and #18, respectively.

ture dependence is similar in form to that of the random sequences [see Fig. 2(b)]. In particular, there is a temperature T_k below which an Arrhenius-like behavior is observed. From Fig. 5(b) this temperature corresponds to $1/T_k \approx 4.5$. An important feature which distinguishes designed sequences from random sequences is that at T_k, the equilibrium folding transition for designed sequences [Fig. 5(a)] is almost finished, and the chain is mainly in the native conformation; for random sequence, by contrast, at its respective T_k $(1/T_k \approx 8)$ the equilibrium transition is far from complete (see Fig. 2). This difference becomes more striking for longer sequences, e.g., for 48-mers, as shown below. This suggests that designed sequences can fold fast at the conditions when their native state is thermodynamically stable.

The fact that the kinetics become Arrhenius-like at the end of the equilibrium folding transition for the designed sequences suggests that below this temperature the energetic and entropic contributions to the free energy barrier for folding do not change with temperature. Such a behavior is in qualitative disagreement with the predictions of Ref. 26. According to Ref. 26, the glass transition is predicted to occur at T_c (T_g in their notation), i.e., at the temperature at which freezing takes place for random sequences having the same amino acid composition (i.e., same E_{av} and σ) as the designed sequences. The value T_c for the present systems was found to be in the range between 8 and 10 (see Fig. 2). No signature of a glass transition is seen in Fig. 5(b) at temperatures close to T_c. Instead, the kinetics of folding appears to be simply related to the thermodynamics in that it becomes Arrhenius-like at the end of equilibrium folding transition. Figure 5(b) also shows the inverse temperature dependence of the median time [squares in Fig. 5(b)]. Comparison of the median time with MFPT [circles in Fig. 5(b)] suggests that the median time is close to MFPT $\cdot \ln 2$ above T_k. This implies (see above) that the folding kinetics can be essentially represented by a single exponential above T_k. However, below T_k a strong deviation of the median time from MFPT $\cdot \ln 2$ can be seen in Fig. 5(b). The deviation from one-exponential behavior can be seen on Fig. 4 at low T as circles (representing MFPT $\cdot \ln 2$) which deviates from the squares of the same color (different grayscale correspond to different sequences on Figs. 2 and 5). This shows that the folding kinetics below T_k is nonexponential for the designed sequences.[3,66] Such behavior contrasts with the single-exponential behavior of random sequences.[36,66]

C. 48-mer

The results discussed so far apply to chains of 27 monomers, which are rather short compared to real proteins. To test whether the behavior found above is general, we simulated a 48-mer. The results obtained are more limited because the folding of longer chains takes more time. We generated ten random sequences for 48-mers and selected the one of them which folded into the lowest energy conformation faster than the other sequences at $T=0.2$. The sequence and the lowest energy conformation are shown on Fig. 6.

Figure 7 shows the inverse temperature dependence of the equilibrium average energy and of the folding rate. Although the average energy as a function of $1/T$ is essentially noncooperative as for the random sequences shown in Figs. 2 and 3 there exists a temperature below which the average energy and the energy barrier for folding (i.e., Arrhenius-like behavior) do not change with temperature. This temperature is equal to $1/T_c \approx 5.5$ from the folding rate and the value is consistent with the behavior of the average energy.

We also designed a sequence with a pronounced energy minimum for the conformation shown in Fig. 6. The results of simulations for this sequence are presented on Fig. 8. As for the 27-mer, the thermodynamic transition is more cooperative than for the random sequence. Again, the kinetics are directly related to the thermodynamics; i.e., neither the equilibrium energy nor the kinetic energy barrier change essentially below the temperature T_k, which can be estimated as $1/T_k \approx 3.2$ from the figure.

IV. PHENOMENOLOGICAL MODEL

Since the results of our simulations are inconsistent with the predictions of the theory developed in Ref. 26, an alternative model is required. We have developed a simple phenomenological model of the folding kinetics for the random sequences. It makes use of the fact that for random sequences the REM is adequate to treat the *thermodynamic*

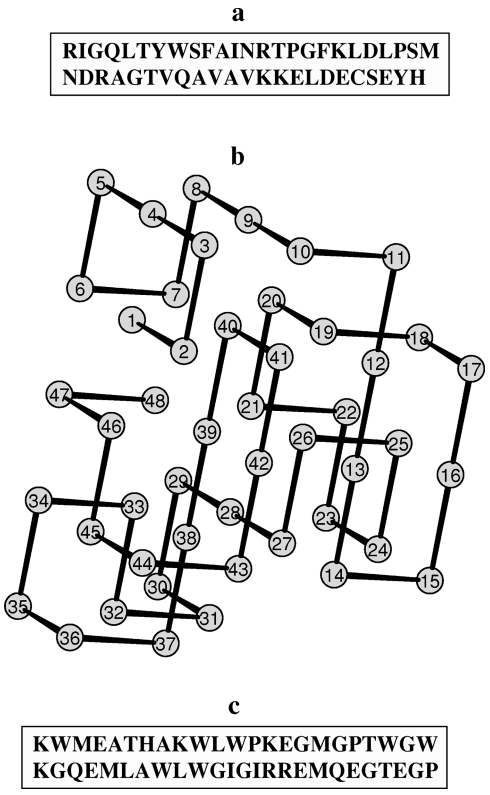

FIG. 6. (a) Random sequence of 48 amino acid, (b) the corresponding native conformation, and (c) sequence designed to minimize relative energy of this conformation.

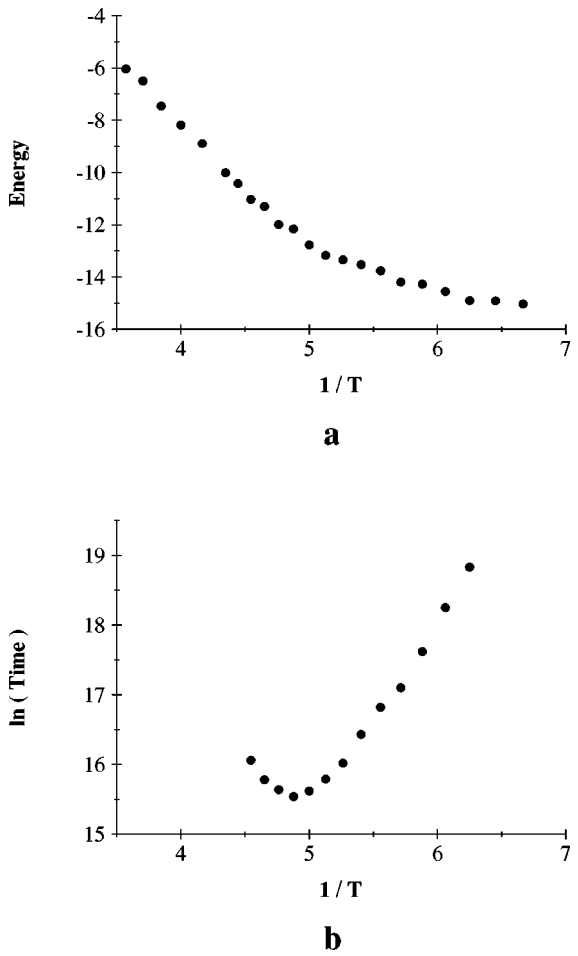

FIG. 7. Inverse temperature dependences (a) of the equilibrium energy E and (b) of the MFPT averaged over 50 runs for random 48-mer [Fig. 6(a)].

properties of the chains, as described in Ref. 1. However, the assumptions to obtain a kinetic model are different (see Discussion).

According to the REM, the density of states $n(E)$ is Gaussian;

$$n(E)=\frac{\Omega}{(2\pi\Sigma^2)^{1/2}}\exp\left[-\frac{(E-E_{\mathrm{av}})^2}{2\Sigma^2}\right]. \tag{4}$$

Here the average energy over all conformations is denoted as E_{av} and $\Sigma^2=2N\sigma^2$ is the standard variance of the energy of the chain. From Eq. (4), it follows that the entropy, ($k_B=1$), is

$$S(E)=\ln n(E)=\ln\Omega-\frac{(E-E_{\mathrm{av}})^2}{2\Sigma^2}, \tag{5}$$

where we have omitted factors that are small in the thermodynamic limit. At the energy, E_c, given by

$$E_c=E_{\mathrm{av}}-\Sigma(2\ln\Omega)^{1/2} \tag{6}$$

the entropy vanishes. This means that the energy spectrum below T_c is sparse; i.e., there are only a few states and they are separated significantly in energy from one another. The ground state energy E_{nat} is only few T_c below E_c.[34] Correspondingly, $(E_c-E_{\mathrm{nat}})/E_c\sim 1/N$; i.e., the deviation of E_{nat} from E_c presents a nonextensive correction so that it suffices to use the approximation $E_c\approx E_{\mathrm{nat}}$ for all estimates of thermodynamic quantities. This can be seen clearly in Fig. 4(a) which gives the density of states for random sequences; i.e., E_c can be estimated from Fig. 4(a) as -5.6 (the lowest energy of conformations which are not structurally similar to the native state, defined as having only 10 native contacts). The native state energy for the sequences whose density of states is shown in Fig. 4(a) is -5.9.

If the temperature is high enough ($T>T_c$), the equilibrium average energy E can be determined from the thermodynamic identity $\frac{dS}{dE}=\frac{1}{T}$. From Eq. (5) we have

$$E=E_{\mathrm{av}}-\frac{\Sigma^2}{T}. \tag{7}$$

This result is valid until the temperature reaches T_c given by

$$T_c=\frac{\Sigma}{(2\ln\Omega)^{1/2}}. \tag{8}$$

At $T<T_c$ the entropy vanishes. The system reaches an energy value where the density of states is low, the free energy is equal to the energy, and the standard thermodynamic relations involving the entropy are no longer valid. From Eq. (7) when the temperature approaches T_c from above, the average energy E reaches its minimal value E_c. This means that, below T_c, E equals to E_c and does not change with temperature in the approximation which neglects nonextensive deviation of E_{nat} from E_c (see above).

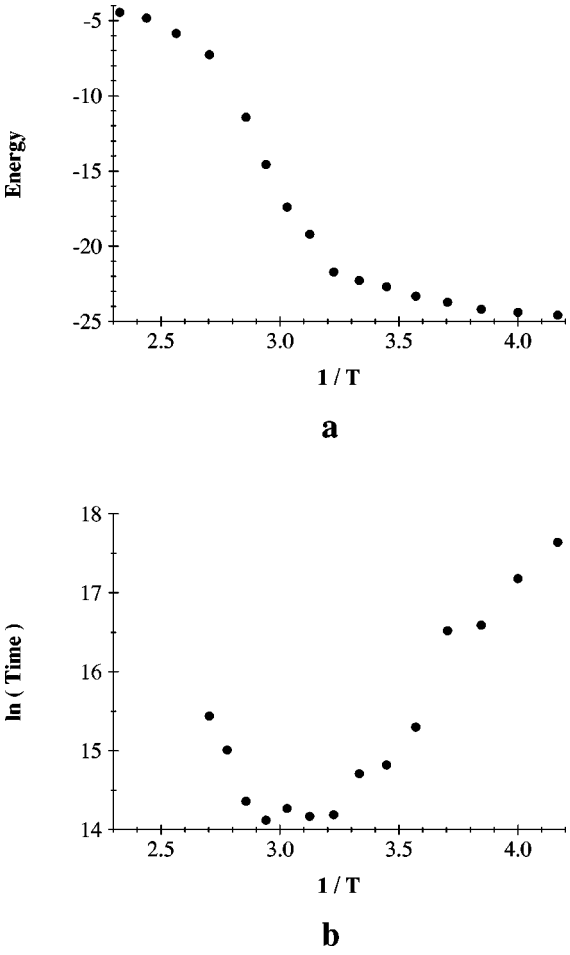

FIG. 8. The same as Fig. 7 but for designed 48-mer [Fig. 6(c)].

To characterize the kinetics of folding of random sequences we need to introduce kinetic postulates since the REM is a model that describes the thermodynamics and not the kinetics, as already stated in the Introduction. The postulates are derived from previous folding simulations and their analysis.[35] They are consistent with the three stage random search mechanism found in simulations for quasi-random sequences in Ref. 35. Since the kinetics of folding to the native conformation for random sequences is a single exponential process, as described above, it can be assumed that folding involves overcoming a free energy barrier, which is the rate-limiting step for the reaction. In the folding simulations, it was found also[35] that the number of conformations which participate in the transition state for folding is much smaller than the total number of semi-compact conformations.

Based on these findings for the folding of random sequences, we make two essential assumptions concerning the kinetics:

(i) The mean energy of the system relaxes relatively rapidly to the equilibrium value determined from the REM given in Eq. (7) for $T>T_c$ and to E_c if $T<T_c$.

(ii) On a slower time scale, the system searches for one of the transition state conformations from which it rapidly folds to the lowest energy (native) conformation. Conformations belonging to the transition state have energy $E^{\#}$ and their number is $\mathcal{N}^{\#}$ so that entropy of the transition state is $S^{\#}=\log \mathcal{N}^{\#}$, which we assume to be temperature-independent.

Postulate (i) which corresponds to the assumption that a small subset of the entire configuration space can be reached rapidly (i.e., the collapsed globule) and that the subsequent search is limited to conformations in this subset, was also suggested in earlier publications.[53,35,67] Postulate (ii) of the kinetic scheme is in the spirit of the kinetic REM discussed in the context of spin glasses by Kopper and Hilhorst[31] and in our earlier publication,[32] as well as in a more recent paper.[33] This postulate was shown to be satisfied for the 27-mer with random sequences (Sali, E.S., and M.K., to be published). It assumes that folding proceeds by an equilibrium mechanism, as do most chemical reactions. Here we give a quantitative phenomenological analysis of the consequences of such a kinetic scheme.

Postulates (i) and (ii) are sufficient to obtain the temperature dependence of the MFPT. At $T>T_c$ the equilibrium energy is given by Eq. (7). Substitution of this equation into Eq. (5) gives the equilibrium entropy

$$S=\ln\,\Omega-\frac{\Sigma^2}{2T^2} \tag{9}$$

and the equilibrium free energy $F=E-TS$.

According to postulate (ii) and the Arrhenius kinetic law, the MFP folding time

$$k_0 t=\exp\left(\frac{F^{\#}-F_{\mathrm{eq}}}{T}\right), \tag{10}$$

where $F^{\#}=E^{\#}-T\log \mathcal{N}^{\#}$ is the free energy barrier and k_0 is the elementary transition rate (without a barrier). It should depend on a number of factors such as the ''connectivity'' of the move set, i.e., how many conformations are connected in one MC step and ''internal viscosity'' which may depend on overall compactness of the chain and reflect steric constraints that may affect the acceptance probability of MC moves. k_0 should be roughly close to the rate of nonspecific collapse of the chain and thus it reflects the nature of polymer dynamics (self-diffusion) of the chain.

Substituting (9) and (7) into the expression (10) for the folding time we obtain

$$\ln\,k_0 t(T)=-\frac{E_c T_c}{2}\left(\frac{1}{T}-\frac{1}{T_c}\right)^2+\frac{E^*}{T}-S^{\#}, \tag{11}$$

where $E^*=E^{\#}-E_c$ is the activation energy.

The entropy of transition state, $S^{\#}$ represents a constant (temperature independent) contribution to the folding time Eq. (11); in our phenomenological theory all temperature-independent terms can be adsorbed into the renormalized elementary rate constant, i.e., we define $\widetilde{k_0}=k_0\,\exp(S^{\#})$ Eq. (11) is valid for $T>T_c$. Below T_c the model is even simpler. We can again use Eq. (10) taking into account that below T_c the entropy vanishes and $E=E_c$. We immediately obtain for $T<T_c$

$$\ln\,\widetilde{k}_0 t(T)=\frac{E^*}{T}. \tag{12}$$

Thus, the analysis predicts the Arrhenius behavior observed in the simulations at low temperatures ($T<T_c$). Above T_c [Eq (11)] there is a parabolic dependence of the MFPT on the inverse temperature. At infinitely high temperature Eq. (11) gives

$$\ln(\tilde{k}_0 t) = -\frac{E_c}{2T_c} = \ln \Omega, \tag{13}$$

that is, the MFPT equals the corrected Levinthal time. This result is as expected, because at infinitely high temperature the search for the native conformation is completely random.

The MFPT as a function of temperature has a minimum at some temperature T_{opt} which can be determined from the Eq. (11). The result is

$$T_{\mathrm{opt}}^{-1} = T_c^{-1}\left(1 + \frac{E^*}{E_c}\right) \tag{14}$$

and the folding time at this temperature is

$$\ln[\tilde{k}_0 t(T_{\mathrm{opt}})] = \frac{E^*}{T_c}\left(1 + \frac{E^*}{2E_c}\right). \tag{15}$$

This should be compared to the folding time at T_c given by

$$\ln[\tilde{k}_0 t(T_c)] = \frac{E^*}{T_c}. \tag{16}$$

The fact that $T_{\mathrm{opt}} \approx T_c$ suggests that $E^* \ll E_c$. This implies that $E^{\#} \approx E_c$, i.e., transition state(s) are closer to the bottom part of the energy spectrum.

The numerical data [Figs. 2(b) and 3(b)] for random sequences were fitted to Eqs. (11) and (12). To do this, we first fitted the Arrhenius portion ($T<T_c$) of the inverse temperature dependence of the MFPT by a straight line [Eq. (12)]. This gives the values of E^* and $\tilde{k}_0$. Then we fitted the high temperature portion ($T>T_c$) by a parabola given by Eq. (11) where E_c is taken to be equal to the lowest energy for a given sequence. Thus, for the high temperature behavior we have only one fitting parameter, T_c. The consistency of the phenomenological theory may be evaluated by comparing T_c obtained as a results of fitting the kinetics with T_c obtained from thermodynamic analysis [Fig. 2(a)]. The results are shown in Fig. 9. It is clear that the phenomenological theory describes the simulation data in a satisfactory manner. In fact the fit is almost perfect for sequence #5 [Fig. 3(b)] but there is a discrepancy at high temperatures, for sequences #1, #4, and #8 which do not include average attractive potential. This deviation is not unexpected because we used a Gaussian density of states [Eq. (4)] with the variance Σ assumed independent of temperature in the derivation. In fact, the variance Σ is proportional to the number of contacts at equilibrium.[16] The number of contacts is insensitive to the temperature at low temperatures. However, for sequences #1, #4, #8 it decreases substantially at high temperatures (data not shown) so that the variance Σ also decreases at high temperatures. The decrease of the variance for these sequences slows down the folding reaction in this regime due to the fact that the number of relevant unfolded conformations to be searched over increases as given by Eq. (5), i.e., the entropic contribution to the free energy becomes important.

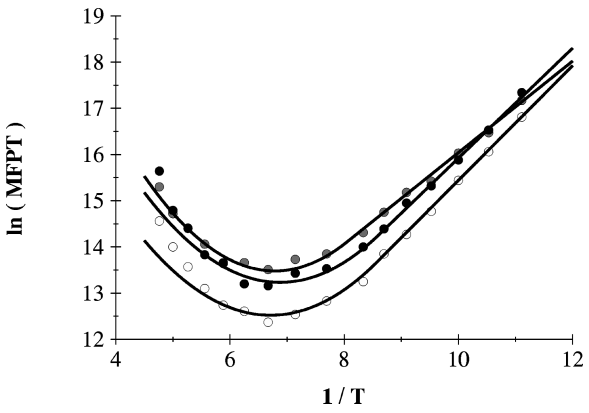

FIG. 9. Fitting of the data from Fig. 2(b) by analytical Eqs. (11) and (12). The parameters of the fitting are as follows: Sequence #1 (black dots): $E_c=-5.92$, $E^*=1.19$, $\ln \tilde{k}_0=-4.2$, $1/T_c=8.60$; Sequence #4 (gray symbols) $E_c=-6.32$, $E^*=0.99$, $\ln \tilde{k}_0=-6.1$, $1/T_c=8.04$; Sequence #8 (white symbols) $E_c=-5.68$, $E^*=1.24$, $\ln \tilde{k}_0=-3.4$, $1/T_c=8.6$.

V. DISCUSSION

The factors which determine the rate of protein folding are of great interest.[34,64,68,69] Most discussions have been motivated by the ''Levinthal paradox'' and possible ways to resolve it. In the last few years, simulations (primarily with lattice models, e.g., Ref. 35), as well as theoretical analyses, have led to the realization that Levinthal paradox is not really a paradox because random search of all possible conformations is not necessary. Instead, the variation in energy of the potential surface of the polypeptide chain plays crucial role. It provides the bias necessary to restrict the search so that folding to the native state can occur in reasonable time. This new perspective on the Levinthal paradox has raised many questions concerning the details of the folding mechanism. A fundamental question concerns the relation between stability, cooperativity and foldability in real proteins. It is clear that the existence of a large energy gap between the native state and states that are significantly different in structure (i.e., so different that the protein would not be active) is a thermodynamic requirement for a stable protein that can perform its biological functions. The magnitude of the required gap obviously depends on the temperature since the dominance of the lowest energy state is determined by the Boltzmann factor; for physiological temperatures a gap of 5 kcal/m leads to a native state population on the order 1000 to one. It was first shown in lattice simulations with a 27-mer chain and interactions derived from a random energy model that the existence of the energy gap is sufficient for the existence of conditions at which protein molecules fold fast into stable native conformation.[34] This result, which supplements and extends previous theoretical work[57,58] has been elaborated in a number of papers that used the energy gap as a criterion for designing fast folding stable sequences.[70,71] The energy separation between the lowest and first excited state can be used as the meaningful energy gap in a fully

compact $3\times3\times3$ cube because the first excited state is generally very different in structure from the native state. For structures that are not fully compact, there can be low energy states that are similar to the native one and so a more general energy gap criterion is appropriate. The one used here [see Eq. (2)] corresponds to that used in matching real proteins by threading through a population of native states.[47]

In two recent papers, the energy gap criterion for fast folding has been called into question. One of these papers[68] neglected the simultaneous requirement of stability and foldability and the other[64] neglected the condition just stated for using the energy difference between the ground and first excited states as the energy gap. It is not our purpose to discuss these two papers here since they will be considered in separate publications. However, their publication makes clear that the energy gap criterion is not understood by everyone working in this field and deserves further discussion.

One important question in applying the energy gap criterion is how to compare the folding rates of different sequences. As has been shown here, as well as in a number of previous studies,[36,3,37] the folding rate of a given sequence is very sensitive to the temperature and exhibits an overall bell shaped curve when the logarithm of the rate is plotted versus the inverse temperature. This might be taken to imply that the folding rates of different sequences should be compared at the optimum temperature (at which folding rate is the fastest) for each. However, in considering the folding rate of a sequence as described by this curve, the stability must not be neglected; i.e., in some cases optimum rate is achieved only at temperature at which sequence is not stable so that the results are not really meaningful as models for proteins. From the present study, it is clear that temperatures corresponding to fastest folding are different for random sequences [with relatively small energy ''gaps,'' see Fig. 4(a)] and designed sequences [with larger ''energy gaps,'' see Fig. 4(b)]. In particular, the fastest folding for random sequences is achieved at lower temperature than the fastest folding for designed sequences, in accord with the theoretical analysis presented in this paper. Indeed, T_c and E_c in Eq. (14) are self-averaging, i.e., their values do not depend on the particular sequence but rather on such averaged characteristics as the amino acid composition.[70,3] The data in Table I and Fig. 2 suggest that this is qualitatively correct; i.e., the value of E_c is constant within $\pm10\%$ for various random sequences given in Table I. The value of the barrier energy $E^{\#}$ (and, thus, $E^*=E^{\#}-E_c$) is different for random and designed sequences. Indeed, since transition state conformations share some structural properties with the native state it is reasonable to assume that their energies follow the Hammond postulate, i.e., transition state energies are lower for the designed sequences. According to the Eq. (14), lower E^* (characteristic of designed sequences) corresponds to higher T_{opt}, which is exactly what is observed in the simulations (compare Figs. 2 and 5). Another feature predicted by the analytical theory is that designed sequences should have $T_{\mathrm{opt}}\approx T_c$ [see Eq. (14)], i.e., that linear part of the bell-curve should begin just below the temperature of fastest folding. Comparison of Figs. 2 and 5 (for 27-mer) and Figs. 7 and 8 (for 48-mer) suggests that this prediction is qualitatively correct.

Our simulations and analysis show that it is possible to find a temperature range at which random sequence folds faster than designed sequence; e.g., at $T=0.14$ $(1/T\approx7)$ the designed 27-mer sequences studied here fold more slowly than random sequences [Compare Fig. 2(a) with Fig. 5(a)]. However, at this high temperature the random sequences are not stable in their native conformations [see Fig. 2(a)]; i.e., although the first passage time is short, the sequences reach their native state and immediately leave it.

It is useful to compare folding rates of sequences under the conditions where their native states are stable. For the random sequences that means that a considerably low temperature must be used than for the designed sequences; i.e., from Fig. 2(a) at the required temperature of 0.09 $(1/T\approx11)$, the folding of the random sequences is about 2 orders of magnitude slower than that of the designed sequences at the temperature 0.2 where they are still stable. However, it should be noted that at $T=0.09$ the designed sequence folds much more slowly than the random sequences. For the 48-mer the corresponding behavior is also observed with the designed sequences folding three orders of magnitude faster at $T=0.35$ $(1/T\approx2.9)$ than random sequences at $T=0.15$ $(1/T\approx6.7)$.

The factor of stability was an essential element in the original formulation of the energy gap criterion for fast folding in Ref. 34 (i.e., folding was studied at a temperature when the sequence is stable). This requirement was neglected, as already mentioned, in a recent publication,[68] so that the conclusions made in this paper are not relevant to the earlier work.

Another way to compare folding rates for different sequences is at the temperature of their respective fastest folding. This was done in a recent paper[71] for random and designed sequences of different lengths in the range 20-100 monomers. The result showed that in this case designed sequences, having a larger ''energy gap,'' on average, folded faster than random sequences, which tend to have smaller energy gaps, This difference became more pronounced as the chain length increased. (The energy gap was defined in Ref. 71 in the same way as Eq. (2) of the present study). However, in line with the above argument, it should be noted here that condition of stability is not satisfied for random sequences at the temperature of their fastest folding.

The study made here of the thermodynamics and kinetics of folding over a wide range of temperatures makes it possible to obtain a deeper understanding of the relevance of different features of protein sequences for fast folding and stability. This also enhances our ability to design sequences which satisfy the requirements of fast folding at a selected folding temperature T. It is clearly advantageous for fast folding to have sequences with the optimal folding temperature T_{opt} close to the actual folding temperature T (a temperature at which sequence is stable). If T_{opt}^0 is the optimal folding temperature of the initial random sequence, the design strategy to generate sequences which fold fast at $T>T_{\mathrm{opt}}^0$ is to make their $T_{\mathrm{opt}}>T_{\mathrm{opt}}^0$. It is clear from Eq. (14) that in this case sequences having higher Σ (and thus higher T_c and E_c) and, most importantly, a lower energy of the native state (a greater energy gap and thus a smaller barrier

E^*, according to Hammond postulate) would emerge as fast folders. Optimization of the energy gap by stabilizing the native state may be also sufficient for stability, which is an important and nontrivial at high T requirement. This is in accord with the original proposal concerning evolution of proteins in archaic high-temperature organisms.[35] At lower temperature, the energy gap requirement for stability is relaxed so that a larger fraction of the sequences have stable native states. Under such conditions fast folding becomes more important because folding rates tend to decrease with temperature as temperature is lowered into actual Arrhenius range. This suggests that additional optimization of sequences to achieve fast folding may be required; i.e., the energy gap criterion by itself may be not sufficient for generating sequences which fold fast at lower temperature.

The present results, as well as earlier studies,[72] suggest how the required optimization can be achieved so that sequences fold fast at $T<T_{\rm opt}^0$. An optimal design strategy would be to adjust sequences to make $T_{\rm opt}$ lower. This can be achieved making T_c lower and E_c higher [see Eq. (14)]. This suggests that in this case optimization may concern factors which determine E_c and T_c, namely, the standard variance of the interaction energies Σ and the average energy of interactions $E_{\rm av}$. Thus, to increase E_c one can increase the average interaction energy (i.e., make monomers more mutually repulsive or less attractive, on average) and/or decrease Σ.

Recently, an evolution-like selection algorithm to generate fast folding sequences was proposed and studied in Ref. 48. In this method sequences are mutated randomly such that only point mutations increasing the folding rate are accepted. This algorithm proved to be efficient in generating fast-folding sequences over a range of temperatures. When it was applied to generate sequences which fold fast at low temperature the resulting sequences had features in accord with the present analysis, i.e., they had a higher average contact energy and a lower dispersion of interaction energies[48] than in the original quasi-random sequences from which the selection began.

Another sequence design algorithm was proposed recently[72] which generated sequences having selected thermal properties, i.e., ones that are stable in a specified temperature range. The analysis presented in this paper suggests that there is a close connection between thermodynamic properties of sequences and kinetics. While the algorithm presented in Ref. 72 selected sequences which are stable in a specified range of temperatures, these sequences had their fastest folding in the same temperature range, which is in line with the findings of the present work. Comparison of sequences generated to fold at high temperature with the ones generated for low temperature showed again that the feature of high-T sequences was that their native states had particularly low energy, i.e., this was effectively an ''energy-gap'' optimization. In contrast, sequences designed to fold and be stable at low T had smaller dispersion of contact interactions, also in accord with the presented analysis.

Although meaningful correlations have been obtained from the above analysis and its applications the results should be considered only as qualitative: A more complete analysis of the relation between stability and fast folding as a function of temperature would require a detailed knowledge of factors affecting E_c and the energy and entropy of the transition state ensemble. In particular the presented phenomenological theory cannot address, in its present form, the important question of chain length dependence of the protein folding time.[71] A consistent analytical theory explaining the power law dependence of folding time on chain length, observed in simulations[71] and supported by qualitative arguments[73] is a matter of future research.

The results of the present simulations and analysis can also be compared with the predictions of the analytical theory developed by Bryngelson and Wolynes[26] as well as more recent extensions and modifications.[51,41,74,75] In this comparison we focus primarily on qualitative aspects of the BW theory and our simulations since quantitative details are sensitive to the specific parameters used in the simulations and analytical calculations and rather arbitrary definitions for ''Levinthal time'' (see above). The BW theory predicts two transitions for designed sequences. One transition is an equilibrium folding transition at temperature T_f; below T_f the native state is thermodynamically stable. At a lower temperature T_c (T_g in the notation of BW) a kinetic glass transition is predicted. The qualitative features of the glass transition as proposed by BW[26,51] are that as the temperature goes down to T_g the folding times reaches a plateau value and no longer changes with temperature and that the kinetics become markedly nonexponential at T_g and below.

Our results do not show two independent transitions. Instead, we observe that both the thermodynamics and kinetics of folding are governed by the same temperature, T_k. This is the temperature below which the equilibrium energy is close to the energy of the native conformation and does not change further with temperature. The simulations results strongly suggest that no qualitative changes in the system properties are expected below T_k. The temperature dependence of the MFPT, as well as of the median time, becomes Arrhenius-like at T_k without any singularities or plateaus at lower temperatures. In this regard the criterion for fast folding, proposed in Ref. 57 can be clarified. It was argued that the ratio ''T_f/T_g'' (a measure of stability gap) could be viewed to determine fast-folding sequences. From the present results it follows that this is not the ratio of folding and ''glass transition'' temperature for the same sequence but rather the ratio of two folding transition temperatures: T_f for the actual sequence and T_g (T_c) which is the folding (''freezing'') transition temperature for a sequence having the same composition as the actual one but which is randomly re-shuffled. Thus this ratio is a measure of sequence optimization.

The temperature dependencies presented in this work (Figs. 2, 3 and 7) rule out, a kinetic glass transition for the system that we studied. Indeed it is impossible to identify a temperature point on the plots of MFPT vs T shown in Figs. 2, 3, 7 with any ''intuitive'' glass transition point. There are only two special temperatures discernible in these plots: $T_{\rm opt}$ and T_k. Obviously, it does not make sense to identify $T_{\rm opt}$, the temperature of fastest folding, with the ''glass transition''; it is equally incorrect to do so for T_k since folding rate at T_k is only a few times slower than the fastest folding rate

(at T_{opt}), both for 27-mer and 48-mer. Further, the kinetics for the random sequences are purely exponential at all temperatures studied, a factor inconsistent with qualitative features of glass transition behavior.[44] Since glass transition does not have a rigorous definition (unlike Ehrenfest definition of phase transitions) only circumstantial evidence for or against it can be obtained. Our results point out to the absence of such a transition in the lattice model that we have studied.

It was claimed earlier[14] that ''glass transition'' behavior had been observed in lattice simulations. Indeed, Fig. 12 of Ref. 14 apparently does show that the Arrenius dependence of the unfolding rate levels off at lower temperature. The discussion of these data in the text of Ref. 14 interprets that behavior as evidence of a glass transition corresponding to that predicted by BW. However, it is explained in the caption to Fig. 12 of Ref. 14 that the form of the Arrenius curve is an artifact of the method used to generate the curve: Simulations were stopped either when the molecule unfolded to a specified degree or when its length reached $T_{max}=1.08\times10^9$ steps. Any simulation at which the chain did not unfold in T_{max} steps was assigned a fixed time of T_{max} steps. Thus the low-temperature plateau of the folding time at T_{max} in the simulations discussed in Ref. 14 represents nothing more than the arbitrary cutoff.

The absence of a ''freezing'' glass transition at T_c in our simulations, in contrast to the theoretical prediction of BW, can be understood by considering the fact that an important assumption in the kinetic model of BW[26] is that transitions occur between conformations with statistically *independent* energies. As a result, a typical conformation with energy close to E_c is ''kinetically connected'' to conformations with energies close to the average energy because the latter are in the vast majority. Correspondingly, the energy barrier E^* to leave any conformation with energy E_c is of order $-E_c$. Thus, the lifetime in any low-energy non-native conformation; according to Ref. 26, is $t\sim\exp(-E_c/T)$. Together with Eq. (16) this gives $t\sim\exp(-E_c/T_c)=\Omega^2$ for $T=T_c$. Since the number of such low-energy conformations (with energy close to E_c) is of order unity, the folding time at $T\leqslant T_c$ is of the order of the time needed to escape from any low-energy non-native conformation that is, the folding time is equal to Ω^2. It was assumed in the subsequent development of the theory that two conformations connected by one kinetic step have finite correlation rather than no correlation at all.[51] This corresponds to the statement that conformations with energy close to E_c are kinetically connected not to conformations with the average energy (which can be taken to be zero) but to conformations with energy αE_c, where α is a parameter which characterizes correlation between two kinetically connected states. This reduces the typical barrier from E_c to $(1-\alpha)E_c$ and leads to the prediction of a faster average rate at T_c than in fully non-correlated case. However, qualitative conclusions, including the prediction of ''glass transition'' of the type suggested in Ref. 26 were not changed in the newer version of the theory introduced in Ref. 51.

It is noteworthy that folding at T_c is much slower in the uncorrelated model of BW than folding at very high temperature where it equals to Ω, the ''Levinthal'' time of exhaustive search without any energetic biases. The uncorrelated model of BW does not indeed provide any energetic biases toward the native conformation. However, the search at low T in the BW model is slower than a simple exhaustive ''Levinthal'' search because it also requires overcoming the energy barriers. It was mentioned in Ref. 26 that the theory presented there may describe kinetics of Monte-Carlo folding simulations. The assumption of the independence of the energies of one-step-connected states is equivalent to a simulation that generates a totally new conformation at each step. At very high temperature each such attempt will be accepted and the system would effectively perform an exhaustive search for the native conformation. The time required for that is equal to Levinthal time Ω. However, at lower temperature, an additional factor slows down folding in the model of BW compared to the unbiased exhaustive search. It is clear that when low energy (E_c) is reached (at $T<T_c$), any randomly generated conformation will have a much higher ($\sim NkT_c$) energy and all such moves except exponentially rare ones will be rejected. This makes the search at T_c inefficient in the model of BW and leads to a folding time Ω^2, much greater than Levinthal time. Decreasing of temperature below T_c gives rise to a further dramatic slowing down since energetic barriers to escape low energy conformations by random search are very high, are of the order of NkT_c.

In fact, BW obtained Ω rather than Ω^2 as the folding time at all $T\leqslant T_c$, (see Fig. 3 of Ref. 26). The independence of the folding rate on the temperature in the BW calculations suggests that there is no energy barrier at low temperatures.[76] This difference between a simple analysis and the BW results appears to be due to the fact that folding time at $T\leqslant T_c$ was obtained by BW as a result of averaging of the folding rate over the ensemble of ''sequences'' (in the context of proteinlike heteropolymers, each ''sequence'' represents a realization of a quenched distribution of energies of states in the dynamic REM model of BW). In Appendix B we explain why the actual folding rate with which the majority of sequences fold in the BW model is very different from the rate averaged over sequences and show how the conclusion about zero energy barriers at low temperature appears as an artifact of averaging the rate over sequences.

In a more realistic model, conformations connected by one kinetic step are expected to be very similar in structure and, therefore, have similar energies. In particular, conformations connected by one Monte-Carlo step to a conformation that has an energy close to E_c have energies close to E_c as well (i.e., for a single MC step the energies differ by a nonextensive value). We conclude that for more realistic models where energies are related along MC trajectory no kinetic freezing transition temperature at which the folding rate drops significantly below the value predicted by the Arrhenius law is expected and none is found in simulations we have presented.

A caveat that always has to be kept in mind comparing theory and simulations is that simulations are restricted to finite size (and relatively small) systems while analytical theory strictly applies to large (infinite) systems. However, typically the signs of phase transitions are clearly discernible

in the temperature dependencies of different quantities obtained from simulations (e.g., sigmoidal or cusp-like curves) which get sharper as the system size increases (dimensional scaling). This is the case, for example, when the thermodynamic theory of proteinlike heteropolymers is compared with simulations: the predictions of mean-field theory for the thermodynamic character of the folding transition in random sequences[10,23,24] and designed ones[57,20,21] is in qualitative agreement with simulation data (see Figs. 2, 5, 7) and results of exhaustive enumeration of short chains.[24] By contrast, our kinetic data do not provide any qualitative indications of a potential (in the limit of large system) glass transition.

It is also important to consider the relevance to proteins of the phenomenological theory and simulations presented here as well as other approaches discussed in this work. Proteins are relatively small (compared to macroscopic systems). It has been suggested[77] that lattice 27-mers correspond to helical proteins composed of 60 residues (hence 48-mers correspond to $\sim$100 amino acid proteins). If this is the case, our study covers the most relevant range of lengths, particularly for proteins that have been studied in folding experiments. Even with the caveat concerning the finite size of the model chains, our results suggest that the concept of a glass transition may be not relevant for understanding the folding of real proteins.

The issue of what features distinguish folding sequences from nonfolding one has been a matter of considerable current study and interest.[34,57,14,78,64,68,79] The present work suggests that certain features should be optimized to provide stability and folding at different temperatures. While no glass transition was found in the present study, the essential features of protein folding kinetics do depend strongly on the temperature. In particular, the dependence of folding rate on temperature is non-monotonic at higher temperatures and exhibits classical Arrhenius behavior at lower temperatures. These findings are supported by recent experiments of Baker and coworkers,[80] who showed that folding rate indeed exhibits Arrhenius behavior when corrected for the temperature dependence of stability. Further, the obtained results are rationalized by a simple analytical theory which also points out what features of sequences may be responsible for stability and fast folding at different temperatures. It will be interesting to examine homologous proteins from organisms which live at different thermal environment to see whether these differences are manifest in their sequences.

ACKNOWLEDGMENTS

This work was supported by NIH Grant No. GM52126 (to E.S.) and NSF (to M.K.) and Jane Coffin Childs Memorial Fund and NIH (to A.S.).

APPENDIX A: THE RANDOM ENERGY MODEL

The random energy model was introduced by Derrida[22] as a simplest nontrivial model of spin glasses. Bryngelson and Wolynes[10] postulated and Shakhnovich and Gutin[16,81] showed for the microscopic model of heteropolymer (in the mean-field replica theory) that thermodynamic properties of random heteropolymers can be described by the REM. There are a number of good accounts of the REM in the literature (see, e.g., Ref. 82). Here we give a brief description of the model (in a form that is slightly generalized and adapted for the case of heteropolymers) to make clear the results that are used in the present study (see also Ref. 1).

The REM was defined by Derrida as phenomenological model based on two postulates:

(1) The system has $\Omega=\gamma^N$ microstates.

(2) The energies of the microstates can be treated as independent random variables with a Gaussian distribution so that density of states $n(E)$ as a function of the energy E obeys the equation

$$n(E)=\Omega\ \exp\left[-\frac{(E-E_{\mathrm{av}})^2}{2N\sigma^2}\right]. \tag{17}$$

Here N is the total number of monomers, γ is the number of conformations per monomer, E_{av} is energy averaged over all conformations, σ is standard variance of interaction energies. Introducing normalized quantities per monomer $e=E/N$ and $e_{\mathrm{av}}=E_{\mathrm{av}}/N$ one can rewrite for the density of states:

$$n(E)=\exp\left[-N\left(\log(\gamma)-\frac{(e-e_{\mathrm{av}})^2}{2\sigma^2}\right]\right. \tag{18}$$

from which one can see immediately that at the critical value of energy $e_c=E_c/N=e_{\mathrm{av}}-(2\log\gamma\sigma^2)^{1/2}$ the system ''runs out of states.'' Specifically, the density of states for $E>E_c$ is very high so that in every interval of energy above E_c many states (conformations) can be found. In contrast at $E<E_c$ the density of states is very sparse and it is unlikely to find a conformation in any *specific* small interval of energies below E_c. Further, it can be shown that the total number of conformations with energy below E_c is ~ 1 - a negligible small fraction of the total number of conformations.

Now consider how energy of a system changes when temperature is varied. As temperature decreases the energy decreases also for temperatures above T_c at which the energy reaches e_c, the lower limit of the dense part of the density of states, e_c. From then on energy can decrease only slightly because, as will be seen shortly, the lowest energy conformation differs from e_c by small (vanishing as N grows) amount. Therefore at $T<T_c$ $E\approx E_c$.

The dependence of the energy on the temperature at $T>T_c$ can be found from the well-known thermodynamic relation

$$\frac{\partial S}{\partial E}=\frac{1}{T}=\frac{\partial[\log(n(E))]}{\partial E} \tag{19}$$

from where we have

$$E=E_{\mathrm{av}}-\frac{N\sigma^2}{T} \tag{20}$$

at $T>T_c$ and

$$E=E_c \tag{21}$$

at $T\leqslant T_c$ with $T_c=\sigma/[2\log(\gamma)]^{1/2}$

Linearization of the last equation around e_c gives the probability of finding a conformation with energy (per monomer) $e<e_c$

$$p(E)=\exp[-N(e-e_c)/T_c] \tag{22}$$

(see also Appendix to Ref. 34). It is clear that it is extremely unlikely to find a conformation in a random heteropolymer with energy (per monomer) much lower than e_c, i.e., the typical value of energy of the global minimum conformation is $\sim T_c/N$ below e_c. This means that the ''gap'' (deviation of the global minimum conformation from e_c) is small in random sequences, much smaller than e_c itself, so that $e_{\min}=e_c+O(1/N)$.

The dependence of the energy on the inverse temperature given in Fig. 2 is fully consistent with the description given by Eqs. (20), (21). E_c and T_c can be determined from this plot as temperature and energy at which the dependence comes to plateau. Having determined these values $\log(\gamma)$ can be determined and finally one gets after simple algebra

$$\Omega=\gamma^N=\exp\left(-\frac{E_c-E_{\rm av}}{2T_c}\right); \tag{23}$$

this is Eq. (3) of the text.

APPENDIX B: ON THE PROCEDURE OF AVERAGING IN KINETICS CALCULATIONS

Averaging over realizations of disordered systems is a very delicate procedure that requires considerable care. In the context of our analysis each realization corresponds to an individual protein sequence. The difficulty with averaging over all sequences lies in the fact that in the ensemble of all sequences some properties may vary widely from sequence to sequence. This raises the question as to the meaning of a property averaged over all sequences. To obtain meaningful results in averaging over all possible sequences one can average only so-called ''self-averaging'' quantities.[83] Average values of such quantities are close to most probable ones and their variation from sequence to sequence is small (usually in the limit of large systems). In this case a typical representative of the ensemble of sequences will have a value of the property that is close to the calculated average, i.e., the average value will be characteristic of the vast majority of sequences.

In the kinetic calculations of BW Eqs. (150) and (28) of Ref. 26 state that at low temperature all sequences fold with the same rate, equal to the average rate. The folding rate in this regime depends *exponentially* on the barrier to escape from misfolded traps [see Eqs. (21)–(28) of Ref. 26]. This implies that folding rate is not self-averaging.

In order to see that consider first a simple example of the ''ensemble'' of 20^N ($N\gg 1$) sequences such that *one* sequence folds ''instantly,'' barrier free with the rate R_0, one sequence folds with high barrier $B_{\rm high}$, having folding rate is $R_{\rm slow}=R_0\ \exp(-B_{\rm high}/T)$ and the remaining *vast* majority of sequences (i.e., 20^N-2) fold with some ''typical'' barrier $B_0<B_{\rm high}$, having the rate $R_{\rm typical}=R_0\ \exp(-B_0/T)$. The rate averaged over the ensemble of sequences then would be

$$\overline{R(T)}=20^{-N}R_0+20^{-N}R_0\ \exp(-B_{\rm high}/T)+(1-2\cdot 20^{-N})R_0\ \exp(-B_0/T). \tag{24}$$

It is easy to see that at high T the average rate will be close to $R_{\rm typical}$, the rate with which the majority of sequences fold. However, at low $T<B_0/(N\log 20)$ the first term in Eq. (24) dominates, and the average folding rate becomes

$$\overline{R(T)}_{T\to 0}=20^{-N}R_0. \tag{25}$$

The folding time calculated as $t=1/\overline{R(T)}$ turns out to be temperature independent and equal to the ''Levinthal time'' $=20^N R_0^{-1}$ at low temperature. It is clear that this result is an artifact of the averaging of the rate which, at low temperatures is dominated by one ''superfast''-folding sequence rather than the majority of sequences in the ensemble. We note that it is equally inappropriate to average folding time: In this case at low temperature the average time will be dominated by one sequence having ''superslow'' folding with the barrier $B_{\rm high}$. It is also clear that the correct way to proceed is to calculate average *barrier* $\overline{B}$ (i.e., *logarithm* of the rate). In this case folding rate defined as $R_0\ \exp(-\overline{B}/T)$ will coincide with the rate with which the majority of sequences fold (cf. the correct averaging over quenched disorder in thermodynamics which requires averaging of free energy, i.e., *logarithm* of the partition function, which is a sum of Boltzmann exponentials[83]). We note that the phenomenological theory presented in this paper estimates the folding rate with which the majority of sequences fold by estimating a ''typical'' free energy barrier rather than average folding rate.

Technically, BW calculate the average rate in their theory in the following way [e.g., see Eqs. (22), (28), (123), (147)–(150)] of Ref. 26:

$$\overline{R}=\int_0^\infty P(B)\exp(-B/T), \tag{26}$$

where $P(B)$ is the probability distribution for the barrier $B=E(T)-E^{\#}$. $E(T)$ is the equilibrium (at a given temperature) energy. As in the previous, oversimplified, example, at $T\to 0$ the integral in Eq. (26) is dominated by small $B\leqslant T$, i.e., by very few ''sequences'' that have barrier free folding to the native state. At low temperature $T<T_c$ $E(T)=E_c$. It follows that $\ln\overline{R}\to\ln P(E_c-E^{\#}=0)$ as $T\to 0$. $P(E_c-E^{\#}=0)$ represents the probability that a given realization of the landscape (''sequence'') has a barrier free transition to the native state. BW assumed the distribution of barriers to be Gaussian, like the distribution of states in the REM. In this case, for uncorrelated state $P(E_c-E^{\#}=0)=1/\Omega$. Therefore $\overline{R}\sim 1/\Omega$, i.e., temperature-independent folding rate at $T\leqslant T_c$, with ''Levinthal'' folding time $1/\overline{R}=\Omega$, the result shown in Fig. 3 of Ref. 26.

In the REM kinetic calculation the averaging of the rate over the ''ensemble of sequences'' is equivalent to averaging of the rate of transition between all states and the native state. Clearly, one state is just the native state itself and the rate of ''folding'' from the native state to itself is just instant,

''microscopic'' one. It is this ''fast-folding'' but rare event that gives the dominant contribution to the *average* folding rate at low temperature in the kinetic REM model.

[1] M. Karplus and E. Shakhnovich, *Protein Folding* (Freeman, New York, 1992), Chap. 4, pp. 127–196.
[2] R. Zwanzig, A. Szabo, and B. Bagchi, Proc. Natl. Acad. Sci. USA **89**, 20 (1992).
[3] V. Abkevich, A. Gutin, and E. Shakhnovich, J. Chem. Phys. **101**, 6052 (1994).
[4] J. T. Ngo, J. Marks, and M. Karplus, in *Protein Folding Problem & Tertiary Structure Prediction* (Birkhauser, Boston, 1994).
[5] V. Abkevich, A. Gutin, and E. Shakhnovich, J. Mol. Biol. **252**, 460 (1995).
[6] M. Karplus and D. Weaver, Nature (London) **160**, 404 (1976).
[7] P. Kim and R. Baldwin, Annu. Rev. Biochem. **51**, 459 (1982).
[8] K. Dill, K. Fiebig, and H. S. Chan, Proc. Natl. Acad. Sci. USA **90**, 1942 (1993).
[9] Y. Ueda, H. Taketomi, and N. Go, Int. J. Peptide Prot. Res. **7**, 445 (1975).
[10] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).
[11] S. Govindarajan and R. Goldstein, Biopolymers **36**, 43 (1995).
[12] A. Dinner, A. Sali, and M. Karplus, Proc. Natl. Acad. Sci. USA **93**, 8356 (1996).
[13] R. Baldwin, Nature (London) **369**, 183 (1994).
[14] J. Bryngelson, J. N. Onuchic, N. D. Socci, and P. Wolynes, Proteins: Struct. Funct. and Genetics **21**, 167 (1995).
[15] K. Dill *et al.*, Protein Sci. **4**, 561 (1995).
[16] E. I. Shakhnovich and A. M. Gutin, Biophys. Chem. **34**, 187 (1989).
[17] T. Garel and H. Orland, Europhys. Lett. **6**, 307 (1988).
[18] E. I. Shakhnovich and A. M. Gutin, J. Phys. A **22**, 1647 (1989).
[19] C. Sfatos, A. M. Gutin, and E. I. Shakhnovich, Phys. Rev. E **48**, 465 (1993).
[20] S. Ramanathan and E. Shakhnovich, Phys. Rev. E **50**, 1303 (1994).
[21] V. Pande, A. Yu Grosberg, and T. Tanaka, Phys. Rev. E **51**, 3381 (1995).
[22] B. Derrida, Phys. Rev. B **24**, 2613 (1981).
[23] E. I. Shakhnovich and A. M. Gutin, J. Chem. Phys. **93**, 5967 (1990).
[24] A. Dinner, A. Sali, M. Karplus, and E. Shakhnovich, J. Chem. Phys. **101**, 1444 (1994).
[25] E. I. Shakhnovich and A. M. Gutin, Nature (London) **346**, 773 (1990).
[26] J. D. Bryngelson and P. Wolynes, J. Phys. Chem. **93**, 6902 (1989).
[27] H. Frauenfelder and P. Wolynes, Phys. Today **47**, 58 (1994).
[28] P. Wolynes, J. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995).
[29] N. Go and H. Abe, Biopolymers **20**, 991 (1981).
[30] C. De Dominicis, H. Orland, and F. Lainee, J. Phys. (France) Lett. **46**, L463 (1985).
[31] G. Kopper and H. Hilhorst, Europhys. Lett. **3**, 1213 (1987).
[32] E. I. Shakhnovich and A. M. Gutin, Europhys. Lett. **9**, 569 (1989).
[33] J. G. Saven, J. Wang, and P. Wolynes, J. Chem. Phys. **101**, 11037 (1994).
[34] A. Sali, E. I. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[35] A. Sali, E. I. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).
[36] M. Karplus, Caflish, A. Sali, and E. I. Shakhnovich, in *Modelling of Biomolecular Structures and Mechanisms* (Kluwer Academic, Dordrecht, 1994).
[37] N. Socci and J. Onuchic, J. Chem. Phys. **101**, 1519 (1994).
[38] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, J. Chem. Phys. **101**, 8246 (1994).
[39] H. S. Chan and K. A. Dill, J. Chem. Phys. **100**, 9238 (1994).
[40] Z. Guo and D. Thirumalai, Biopolymers **35**, 137 (1995).
[41] N. D. Socci, J. N. Onuchic, and P. Wolynes, J. Chem. Phys. **104**, 5860 (1996).
[42] E. I. Shakhnovich, G. M. Farztdinov, A. M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).
[43] S. Myazawa and R. Jernigan, Macromolecules **18**, 534 (1985).
[44] C. A. Angell, Science **267**, 1924 (1995).
[45] V. Abkevich, A. Gutin, and E. Shakhnovich, Biochemistry **33**, 10026 (1994).
[46] P. H. Verdier, J. Chem. Phys. **59**, 6119 (1973).
[47] J. U. Bowie, R. Luthy, and D. Eisenberg, Science **253**, 164 (1991).
[48] A. Gutin, V. Abkevich, and E. Shakhnovich, Proc. Natl. Acad. Sci. USA **92**, 1282 (1995).
[49] V. Abkevich and E. Shakhnovich, Proteins (in press, 1993).
[50] R. Miller, C. Danko, M. J. Fasolka, A. C. Balazs, H. S. Chan, and K. A. Dill, J. Phys. Chem. **96**, 768 (1992).
[51] S. Plotkin, J. Wang, and P. Wolynes, Phys. Rev. E **53**, 6271 (1996).
[52] C. Levinthal, *Mossbauer Spectroscopy of Biological Systems* (University of Illinois Press, Urbana, Illinois, 1969).
[53] K. Dill, Biochemistry **24**, 1501 (1985).
[54] C. Itzykson, H. Orland, and C. De Dominicis, J. Phys. (France) Lett. **46**, L353 (1985).
[55] A. Gutin, V. Abkevich, and E. Shakhnovich, Biochemistry **34**, 3066 (1995).
[56] M. F. Sykes, J. Chem. Phys. **39**, 410 (1963).
[57] R. Goldstein, Z. A. Luthey-Schulten, and P. Wolynes, Proc. Natl. Acad. Sci. USA **89**, 4918 (1992).
[58] E. Shakhnovich and A. Gutin, Proc. Natl. Acad. Sci. USA **90**, 7195 (1993).
[59] M.-H. Hao and H. Scheraga, J. Phys. Chem. **98**, 4940 (1994).
[60] M.-H. Hao and H. Scheraga, J. Phys. Chem. **98**, 9882 (1994).
[61] S. White and R. Jacobs, Biophys. J. **57**, 911 (1990).
[62] O. B. Ptitsyn, J. Mol. Struct.: THEOCHEM **123**, 45 (1985).
[63] V. Pande, A. Yu Grosberg, and T. Tanaka, Proc. Natl. Acad. Sci. USA **91**, 12972 (1994).
[64] D. Klimov and D. Thirumalai, Phys. Rev. Lett. **76**, 4070 (1996).
[65] D. Klimov and D. Thirumalai, Proteins **26**, 411 (1996).
[66] C. Sfatos, V. Abkevich, A. Gutin, and E. Shakhnovich, Biochemistry **35**, 334 (1996).
[67] C. Camacho and D. Thirumalai, Phys. Rev. Lett. **71**, 2505 (1993).
[68] R. Unger and J. Moult, J. Mol. Biol. **250**, 988 (1996).
[69] O. Galzitskaya and A. Finkelstein, Protein Eng. **8**, 883 (1995).
[70] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).
[71] A. Gutin, V. Abkevich, and E. Shakhnovich, Phys. Rev. Lett. **77**, 5433 (1996).
[72] M. Morrissey and E. Shakhnovich, Folding & Design **1**, 391 (1996).
[73] D. Thirumalai, J. Phys. I **5**, 1457 (1995).
[74] S. Plotkin, J. Wang, and P. Wolynes, J. Chem. Phys. **106**, 2932 (1996).
[75] S. Plotkin, J. Wang, and P. Wolynes, J. Phys. (France) **7**, 395 (1997).
[76] BW explain the temperature independence of the folding time by the approximate character of their model that treats the density of states as a continuum and neglects the transitions from deep traps to other states over nonextensive (in N) barriers. In other words, BW state that at low T, transitions occur over zero, or very small (nonextensive in N) energy barriers, in contrast to the situation at $T>T_c$, when transitions in the BW model occur over larger (extensive in N) barriers.
[77] J. Onuchic, P. Wolynes, Z. Luthey-Schulten, and N. Socci, Proc. Natl. Acad. Sci. USA **92**, 3626 (1995).
[78] M. Karplus and A. Sali, Curr. Opin. Struct. Biol. **5**, 58 (1995).
[79] E. I. Shakhnovich, Curr. Opin. Struct. Biol. **7**, 29 (1997).
[80] L. M. Scalley and D. Baker, Proc. Natl. Acad. Sci. USA **94**, 10636 (1997).
[81] E. I. Shakhnovich and A. M. Gutin, Europhys. Lett. **8**, 327 (1989).
[82] K. Binder and A. Young, Rev. Mod. Phys. **58**, 801 (1986).
[83] M. Mezard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1988).