Master equation approach to finding the rate-limiting steps in biopolymer folding

Wenbing Zhang and Shi-Jie Chen[a)]
Department of Physics and Astronomy and Department of Biochemistry, University of Missouri, Columbia, Missouri 65211

(Received 18 September 2002; accepted 22 November 2002)

A master equation approach is developed to find the rate-limiting steps in biopolymer folding, where the folding kinetics is described as a linear combination of basic kinetic modes determined from the eigenvalues and eigenvectors of the rate matrix. Because the passage of a rate-limiting step is intrinsically related to the folding speed, it is possible to probe and to identify the rate-limiting steps through the folding from different unfolded initial conformations. In a master equation approach, slow and fast folding speeds are directly correlated to the large and small contributions of the (rate-limiting) slow kinetic modes. Because the contributions from the slow modes can be computed from the corresponding eigenvectors, the rate-limiting steps can be identified from the eigenvectors of the slow modes. Our rate-limiting searching method has been tested for a simplified hairpin folding kinetics model, and it may provide a general transition state searching method for biopolymer folding. © *2003 American Institute of Physics.* [DOI: 10.1063/1.1538596]

I. INTRODUCTION

A central problem in biopolymer folding kinetics is to understand the transition states, or, the rate-limiting steps. Efficient computational methods have been developed to search for the transition states (as saddle points on the potential energy surface)[1–4] for small molecule chemical kinetics, where well defined molecular structures of the reactant, product, and the transition state can be specified. For a macromolecule, like a protein and an RNA molecule, however, the molecule moves over a large region on the energy landscape, and there are many parallel pathways that lead to the final equilibrium state of the system. As a result, the transition state consists of an ensemble of rate-determining chain conformations. Such rate-determining chain conformations often contain certain critical structural elements whose formation or disruption would be rate-limiting. In this paper we present a method to search for the rate limiting steps.

To identify the rate-limiting steps has been a long standing problem in biopolymer folding kinetics. Several remarkable computational methods have been developed to search for the transition states,[5–29] mostly based on Monte Carlo and molecular dynamics simulations for the folding from randomly chosen unfolded initial conformations. The approach reported in this paper is motivated by several recent master equation based statistical mechanical models for protein[30–39] and RNA[40,41] folding. The method reported here can account for the complete ensemble of chain conformations, and thus enables rigorous and deterministic searching for the rate-limiting steps.

II. MASTER EQUATION APPROACH TO THE FOLDING KINETICS

In a master equation description, the kinetics for the fractional population (or the probability) $p_i(t)$ for the ith state ($i=0,1,..,\Omega-1$, where Ω is the total number of chain conformations) is described as the difference between the rates for transitions entering and leaving the state

$$dp_i(t)/dt=\sum_{j=0}^{\Omega-1}[k_{j\rightarrow i}p_j(t)-k_{i\rightarrow j}p_i(t)],$$

where $k_{j\rightarrow i}$ and $k_{i\rightarrow j}$ are the rate constants for the respective transitions. The above master equation has an equivalent matrix form: $d\mathbf{p}(t)/dt=\mathbf{M}\cdot\mathbf{p}(t)$, where $\mathbf{p}(t)$ is the fractional populational vector col $(p_0(t),p_1(t),\ldots,p_{\Omega-1}(t))$, $\mathbf{M}$ is the rate matrix defined as $M_{ij}=k_{i\rightarrow j}$ for $i\neq j$ and $M_{ij}=-\Sigma_{l\neq i}k_{il}$ for $i=j$.

For a given initial folding condition c at $t=0$, the master equation solution yields the following populational kinetics $\mathbf{p}(t)$ for $t>0$:

$$\mathbf{p}(t)=\sum_{m=0}^{\Omega-1} C_m^{(c)}\ \mathbf{n}_m\ e^{-\lambda_m t}, \tag{1}$$

where $-\lambda_m$ and $\mathbf{n}_m$ are the mth eigenvalue and eigenvector of the rate matrix $\mathbf{M}$.[42] The eigenmodes are labeled according to the monotonically decreasing order of the eigenvalues $(-\lambda_m)$: $0=\lambda_0<\lambda_1\leqslant\lambda_2\leqslant\ldots\leqslant\lambda_{\Omega-1}$. The eigenvector $\mathbf{n}_0$ for the equilibrium mode $m=0$ gives the equilibrium populational distribution.

According to Eq. 1, the overall folding kinetics is a linear combination of the Ω kinetic modes weighted by the coefficient $C_m^{(c)}$. Therefore, the coefficient $C_m^{(c)}$ represents the contribution from the mth mode to the overall kinetics. Each kinetic mode m in Eq. (1) represents an elementary

[a)]Author to whom correspondence should be addressed; electronic mail: chenshi@missouri.edu

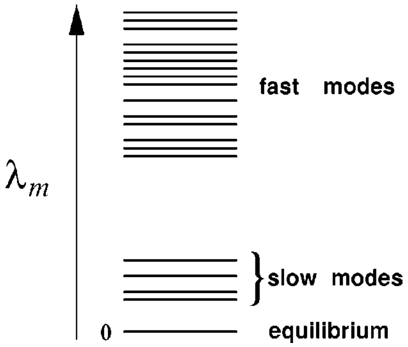

FIG. 1. A large gap between the eigenvalues of the (rate-determining) slow modes and those of the fast modes.

concerted kinetic process (of rate λ_m) in which the populational variation of each state is proportional to the respective component in the eigenvector $\mathbf{n}_m$.

If there exists a large gap between the eigenvalues of a group of slow modes of small λ_m values and the eigenvalues of the fast modes of large λ_m values (see Fig. 1), the overall folding would be rate limited by the slow modes, and the folding reaction would show multiple time scales kinetics: the molecule undergoes a fast initial relaxation due to the fast kinetic modes, followed by a slow rate-determining process due to the slow kinetic modes at a later stage. Especially, if there exists a single distinctive slow kinetic mode m_s, the single slow kinetic mode m_s gives the bottleneck kinetic step and the overall folding time is given by $t_{\rm fold} \sim \lambda_{m_s}^{-1}$.

For an initial state c defined by the populational distribution $\mathbf{p}(0)$ at $t=0$, according to the orthogonality condition of the eigenvectors, the weighting coefficients $C_m^{(c)}$ in Eq. 1 can be determined by the following equation:

$$C_m^{(c)} = \sum_{i=0}^{\Omega-1} p_i(0) n_m^{(i)}/n_0^{(i)},$$

where $n_m^{(i)}$, $n_0^{(i)}$, and $p_i(0)$ are the ith components of vectors $\mathbf{n}_m$, $\mathbf{n}_0$, and $\mathbf{p}(0)$, respectively. Especially, for a folding process starting from a specific conformation c ($c = 0,1,2,\ldots,\Omega-1$) such that the initial populational state $\mathbf{p}(0)$ is given by vector $(0,\ldots,0,$ 1 (the cth component), $0,\ldots 0)$, the weighting coefficient $C_m^{(c)}$ can be conveniently computed from the eigenvectors: $C_m^{(c)} = n_m^{(c)}/n_0^{(c)}$.

III. MODEL AND METHODS

A rate-limiting step can be the formation or breaking of a critical structure (e.g., a specific base pair in RNA folding) such that the folding will proceed quickly once such a critical structure is formed or disrupted. A folding process can involve multiple rate-limiting steps. In the following, we use a single rate-limiting step to illustrate the idea. We will later show that the same approach can be applied to the folding kinetics involving multiple rate-limiting steps. As shown in Fig. 2, for a single rate-limiting process:

(1) if the formation of a critical intrachain contact (a,b) is an (on-pathway) folding rate-limiting step (e.g., the formation of a nucleus in the nucleation process), folding starting from any of the conformations that contain the (a,b) contact would be fast;
(2) if the disruption of a critical intrachain contact (c,d) is an (off-pathway) folding rate-limiting step (e.g., breaking a kinetic trap), folding starting from any of the conformations that contain the (c,d) contact would be slow.

So the on-pathway and off-pathway rate-limiting steps can be identified by searching for the common contacts contained in the fast and slow folding conformations.

Because the rate matrix and the master equation account for the complete ensemble of chain conformations, the eigenvalues and eigenvectors can give the populational kinetics for each and every state for any given initial conformation. We can exactly exhaustively enumerate the possible conformations to find out all the fast and slow folding conformations. However, as we present in the following, such exhaustive information about the fast and slow folding conformations is effectively contained in the eigenvectors for the slow modes.

We denote the (single) slow mode as the m_sth mode, i.e., $\lambda_{m_s} \ll \lambda_m (m \neq 0, m_s)$. The folding speed is directly correlated to the weighting coefficient $C_{m_s}^{(c)}$ in Eq. (1) because of the following reason. The folding reaction starting from the initial conformation c would be *slow* if the process is domi-

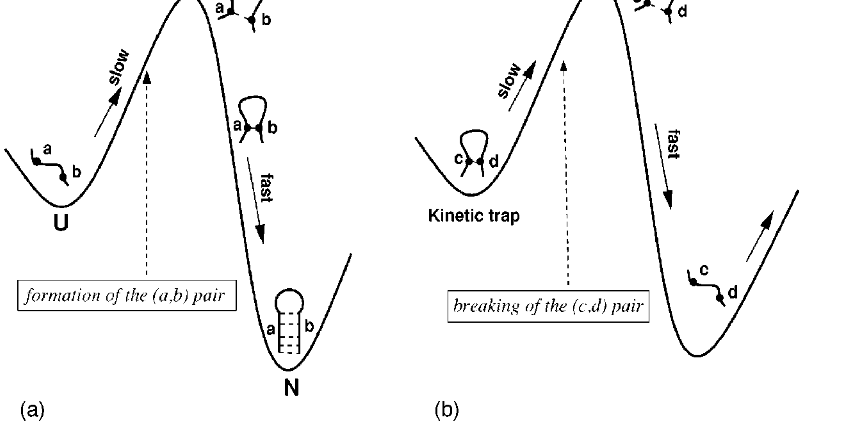

FIG. 2. (a) If the formation of a critical base pair (a,b) is a (on-pathway) rate-limiting step (e.g., for nucleation), folding starting from any of the conformations that contain the (a,b) base pair would be fast, and all conformations that lead to *fast* folding would contain the rate-limiting base pair (a,b) (if there is a single rate-limiting step). (b) If the breaking of a critical base pair (c,d) is a (off-pathway) rate-limiting step (to leave a kinetic trap), all conformations that lead to *slow* folding would contain the rate-limiting base pair (c,d) (if there is a single rate-limiting step).

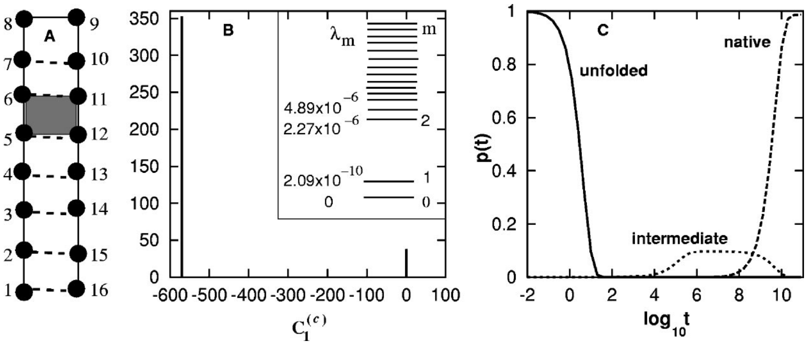

FIG. 3. (A) The native structure of the RNA-like hairpin. The formation of the shaded stack (5,6,11,12) is an (designed) on-pathway rate-limiting step. (B) The eigenvalue distribution (inset) and the histogram of $C_1^{(c)}$ for the slow mode $m_s = 1$. (C) The populational kinetics in the folding process. The kinetic intermediates correspond to the transient accumulation of the conformations before the (5,6,11,12) is formed. In all the populational curve plots presented in this paper, the results for the states of which the fractional populations have never exceeded 10% during the folding process are not shown.

nated by the slow mode m_s, corresponding to a *large* magnitude of the weighting coefficient $C_{m_s}^{(c)}$ in Eq. (1), and the folding would be *fast* if the process is dominated by the fast modes, corresponding to a *small* magnitude of the weighting coefficient $C_{m_s}^{(c)}$. Therefore, for a single rate-limiting step case, all conformations c can be unambiguously classified into two (fast and slow conformational) groups according to the (small and large) values of $|C_{m_s}^{(c)}|$.

Therefore, for a folding reaction starting from the fully unfolded state U, the on- and off-pathway rate-limiting steps can be identified in the following way:

(1) if the fully unfolded state U belongs to the group of the slow folding initial conformation c_s's that have large $|C_{m_s}^{(c_s)}|$ values, there is an on-pathway rate-limiting step, corresponding to the formation of the structure that exists in all the fast folding initial conformation c_f's that have small values of $|C_{m_s}^{(c_f)}|$;

(2) if the fully unfolded state U belongs to the fast folding initial conformation c_f's that have small $|C_{m_s}^{(c_f)}|$ values, there is an off-pathway rate-limiting step, corresponding to the disruption of the (misfolded) structure that exists in all the slow folding initial conformation c_s's that have large values of $|C_{m_s}^{(c_s)}|$.

For a folding process involving multiple rate-limiting steps, as we will show below, there are multiple slow modes, and for each slow mode m_s, the weighting coefficient $|C_{m_s}^{(c)}|$ has a hierarchical (discrete) distribution for different initial folding conformations c, corresponding to the passage of the different (combinations) of rate-limiting steps. Therefore, the chain conformations can be classified into a hierarchy of different subgroups according to the $|C_{m_s}^{(c)}|$ values. To find out *all* the rate-limiting steps, we can apply the above method to *each* of the slow modes m_s. For each slow mode m_s, the $|C_{m_s}^{(c)}|$ spectrum gives on-pathway rate-limiting steps if U belongs to the slowest folding subgroup, and gives off-pathway rate-limiting steps, otherwise.

IV. TEST IN A SIMPLIFIED HAIRPIN FOLDING KINETICS MODEL

The purpose of the calculations presented in the following is to illustrate the principle and to test the above rate-limiting step searching method using a simplified hairpin folding model. Because our purpose here is not for the complete investigation for hairpin folding kinetics *per se*, we choose a minimal model that can have a variety of complex on- and off-pathway rate-limiting steps in order to simplify the analysis and to focus on the test for the method. Our strategy is to design rate-limiting steps by assigning proper entropy and enthalpy parameters for the formation (disruption) of certain rate-determining intra-chain contacts or base stacks (= adjacent stacking base pairs in nucleic acids). We then use the above rate-limiting step searching method to show how to find all the *designed* on- and off-pathway rate-limiting steps.

We assume that the conformational stability is provided by the base stacking force, and we use base stacks to describe a state.[40] We assume an entropic (ΔS_0) and enthalpic (ΔH_0) change of $(\Delta S_0, \Delta H_0) = (-5, -2)$ for the formation of a base stack, unless specified otherwise for certain special base stacks involved in the rate-limiting steps in specific models. All possible hairpin-forming base stacks and loops are allowed to form in the model. Loop entropies are neglected in order to simplify the illustrative calculation. It must be pointed out that the sequence-dependent enthalpy and entropy parameters for the formation (disruption) of the base stacks and small loops have been experimentally measured for RNAs.[43] However, here we use such a rather simplified minimal model for the purpose of ideally simplifying the analysis in order to focus on the analysis for the rate-limiting step searching method.

In our simplified model, there are totally 391 hairpin conformational states for a 16-nt chain, and the native structure is a hairpin with 6 base stacks; see Fig. 3(A). The rate matrix is defined as $k_{i\to j} = e^{-\Delta G_{ij}/k_B T}$ for the formation or the disruption of a base stack and $k_{i\to j} = 0$ otherwise. The kinetic barrier ΔG_{ij} is assumed to be entropic and equal to $-T\Delta S_{ij}$ for the formation of a base stack and ΔG_{ij} is assumed to be enthalpic and equal to $-\Delta H_{ij}$ for the disruption of a base stack,[40] where ΔH_{ij} and ΔS_{ij} are the enthalpic and entropic change for the $i \to j$ transition. In all the models presented in the following, the chain has a melting temperature of $T_m \simeq 0.4$.[44–47] We will study the equilibration (folding) kinetics under folding conditions of temperature $T < T_m$.

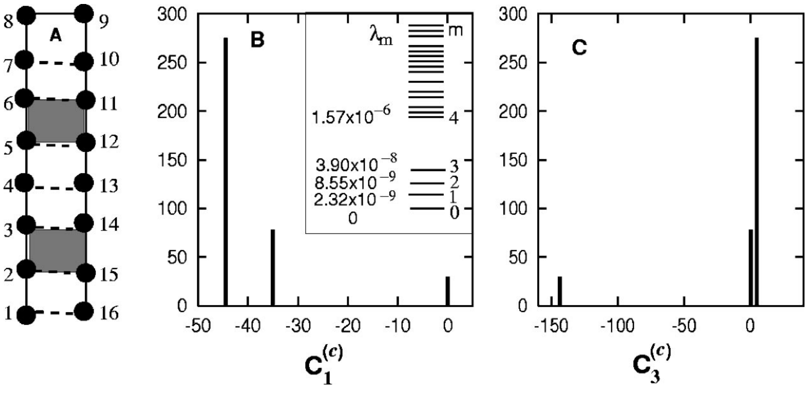

FIG. 4. (A) The native structure of the RNA-like hairpin. The formation of the shaded stacks (2,3,14,15) and (5,6,11,12) are two (designed) on-pathway rate-limiting steps. (B) The eigenvalue distribution (inset) and the histogram of $C_1^{(c)}$ for the slow mode $m_s = 1$. (C) $C_3^{(c)}$ for the slow mode $m_s = 3$.

A. On-pathway rate-limiting step

Model. We design an on-pathway rate-limiting folding process by assigning a large entropic barrier $-\Delta S$ for the formation of the native base stack (5,6,11,12): $(\Delta S, \Delta H) = (-15, -6)$.

Eigenvalue spectrum. We found that the rate matrix for $T = 0.2$ ($< T_m$) has a single slow mode $m_s = 1$ of rate (eigenvalue) $\lambda_1 = 2.08\times 10^{-10}$ [see Fig. 3(B)], causing an overall folding time of $t_{\text{fold}} \sim \lambda_1^{-1}$, which is consistent with the populational kinetics curve shown in Fig. 3(C).

$|C_{m_s}^{(c)}|$**–distribution.** As shown in Fig. 3(B), $C_{m_s}^{(c)}$ for the slow mode $m_s = 1$ shows a 2-value distribution: $C_1^{(c)} \sim 0.0017$ for all the conformations (c) that have the (5,6,11,12) base stack formed, and $C_1^{(c)} \sim -569.6$ for all the conformations without the (5,6,11,12) stack. Therefore, the formation of base stack (5,6,11,12) is the (on-pathway) rate-limiting folding step, and we have found out the (designed) rate-limiting step.

B. Two on-pathway rate-limiting steps

Model. We extend the above model by assuming $(\Delta S, \Delta H) = (-15, -6)$ for both the (5,6,11,12) and (2,3,14,15) base stacks, and so the formation of the native base stacks (5,6,11,12) and (2,3,14,15) are expected to be two rate-limiting steps in series on the folding pathway.

Eigenvalue spectrum. We found that for $T = 0.25$, there are three distinctive slow eigenvalues $2.32\times 10^{-9}, 8.55\times 10^{-9}$ and 3.90×10^{-8}, corresponding to modes $m_s = 1,2,3$, respectively [see Fig. 4(B)].

$|C_{m_s}^{(c)}|$**–distribution.** As shown in Figs. 4(B) and 4(C), $C_{m_s}^{(c)}$ for each slow mode has well separated discrete values. For modes $m_s = 1$ and 3, the group of conformations for the smallest $|C_{m_s}^{(c)}|$ values share the common base stacks (5,6,11,12) and (2,3,14,15), respectively. Therefore, the formation of native base stacks (5,6,11,12) and (2,3,14,15) are two rate-limiting steps. For mode $m_s = 2$, conformations with the smallest $|C_2^{(c)}|$ value have both base stacks formed.

C. Off-pathway rate-limiting process

Model. We assume a large enthalpic barrier $-\Delta H$ for the disruption of a non-native base stack (1,2,5,6): $(\Delta S, \Delta H) = (-15, -6)$, hence conformations with base stack (1,2,5,6) form a kinetic trap [see Fig. 5(A)]. In fact, in the present simplified model, there is only one hairpin conformation that can form the base stack (1,2,5,6).

Eigenvalue spectrum. For the folding kinetics at $T = 0.2$ we found that there exists a single slowest mode $m_s = 1$ of rate (eigenvalue) of $\lambda_1 = 9.39\times 10^{-14}$ [see Fig. 5(B)].

$|C_{m_s}^{(c)}|$**–distribution.** As shown in Fig. 5(B), $C_1^{(c)} \sim -1889$ for the conformation c with the (1,2,5,6) base stack formed, and $C_1^{(c)} \sim 0.00053$ for all other conformations, including the fully unfolded state U; see Fig. 5(B). Therefore, the dominant rate-limiting step is the disruption of the base stack (1,2,5,6).

In fact, the rate constant for the slow mode λ_1 can be estimated from the rate constant for the disruption of the (1,2,5,6) stack: $e^{-\Delta H/k_B T} = e^{-6/0.2} = 9.36\times 10^{-14}$, which agrees exactly with the slowest eigenvalue λ_1 of the rate matrix.

Kinetic partitioning.[48] The rate-limiting slow kinetic modes identified above are intrinsic to the energy landscapes. However, a molecule starting from a given initial state on the

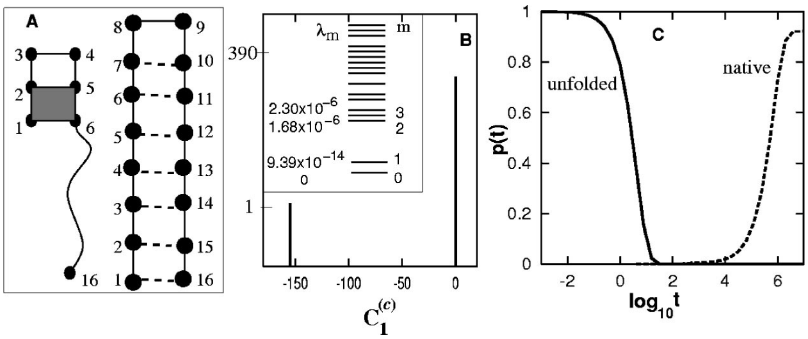

FIG. 5. (A) The native and the (designed) trapped structures with the shaded (1,2,5,6) stack. (B) The eigenvalue distribution (inset) and the histogram of $C_1^{(c)}$ for the slow mode $m_s = 1$. (C) The populational kinetics in the folding process. No kinetic intermediate state is formed because the kinetic trap is avoided in the folding process.

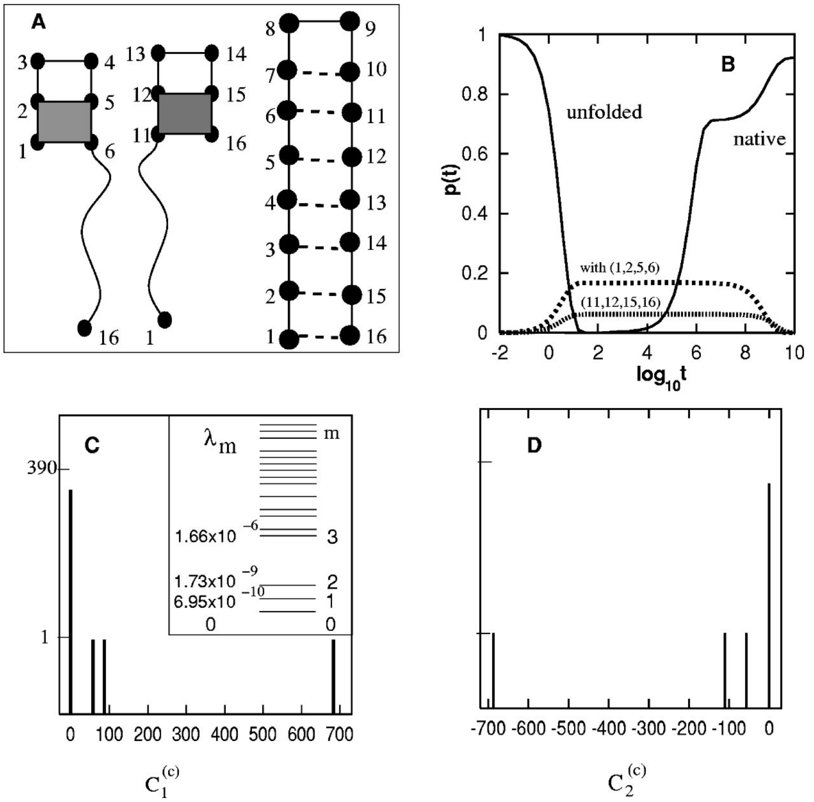

FIG. 6. (A) The native and the (designed) trapped structures with the shaded (1,2,5,6) and (11,12,15,16) stacks. (B) The populational kinetics in the folding process. The kinetic intermediates correspond to the transient accumulation of conformations with (1,2,5,6) and (11,12,15,16) base stacks. The native population shows a biphasic kinetics. (C) The eigenvalue spectrum (inset) and the histogram of $C_1^{(c)}$ for the slow mode $m_s = 1$. (D) The histogram of $C_2^{(c)}$ for the slow mode $m_s = 2$.

energy landscape can fold along different pathways in parallel, many of which do not necessarily pass through the rate-limiting kinetic traps.[48] The overall folding kinetics is not only determined by the existence of the rate-limiting slow kinetic modes and the associated kinetic traps, but is also determined by the likelihood for the molecule to pass through the rate-limiting steps. Mathematically, the likelihood for the molecule to pass through a trapped state r can be characterized by its transient population $p_r(t)$ given by the rth component of the vector equation Eq. (1). Furthermore, for a given starting state c, the peak value of $p_r(t)$ (= the maximum transient accumulation of state r in the kinetic trap) is mainly determined by the coefficients $C_{m_s}^{(c)} n_{m_s}^{(r)}$ in Eq. (1) for the *slow* modes m_s, because the contributions from the *fast* modes decay quickly after initial folding events.

In model C above, as shown in the populational kinetics in Fig. 5(C), the overall folding time $t_{\text{fold}} \sim 6.0 \times 10^5$ is much faster than that predicted from the rate-limiting slow mode which gives $\lambda_1^{-1} \sim 1.0 \times 10^{13}$. Such "anomalous" fast folding kinetics arises from the negligible likelihood for the molecule to pass through the kinetic trap. In this case, c is the initial unfolded state U, the slow mode is $m_s = 1$, and the trapped state r is the one with the (1,2,5,6) base stack, we found the likelihood $C_1^{(U)} n_1^{(r)} = (-0.024)(-0.00645) \ll 1$ is negligible. Therefore, the molecule would most likely fold through the fast kinetic modes $(m \geqslant 2)$, and the rate-limiting slow mode (and the kinetic trap) is avoided. As a result, the overall folding time can be estimated as $\lambda_2^{-1} \sim 6 \times 10^5$ $(\simeq t_{\text{fold}})$.

D. Two off-pathway rate-limiting steps

Model. We assign high enthalpic barriers $-\Delta H$ for the disruption of two non-native base stacks by assuming $(\Delta S, \Delta H) = (-3, -4)$ and $(-4.0, -4.2)$ for the formation of base stacks (1,2,5,6) and (11,12,15,16), respectively [see Fig. 6(A)]. Therefore, mis-formed non-native base stacks (1,2,5,6) and (11,12,15,16) are expected to be two kinetic traps.

Eigenvalue spectrum. The first four eigenvalues for $T = 0.2 (< T_m = 0.4)$ are $0, -6.95 \times 10^{-10}, -1.73 \times 10^{-9}, -1.66 \times 10^{-6}$. Therefore, there are two distinctive slow modes $(m_s = 1,2)$ [see Fig. 6(C)].

$|C_{m_s}^{(c)}|$**–distribution.** Figures 6(C) and 6(D) show that the dominant $C_{m_s}^{(c)}$ values for the two slow modes are 687.0 (for $m_s = 1$) and -687.0 (for $m_s = 2$) for conformations c that contain base stacks (11,12,15,16) and (1,2,5,6), respectively. Therefore, the two rate-limiting slow modes correspond to the breaking of the respective mis-formed base stacks.

Kinetic partitioning. For a folding from the fully unfolded state U, the likelihoods for passing through the two kinetic traps $C_{m_s}^{(U)} n_{m_s}^{(r)}$ are $56 \times 0.00146 \sim 8\%$ and $110 \times 0.00146 \sim 16\%$ for (11,12,15,16) and (1,2,5,6), respectively. Because of the small probability of trapping, the overall folding kinetics becomes biphasic [see Fig. 6(B)]:

(a) $t \sim \lambda_3^{-1} \sim 10^6$: the native population quickly rises to a significant level because the molecule has a large probability to fold quickly through the fast modes $(m \geqslant 3)$ without being trapped;

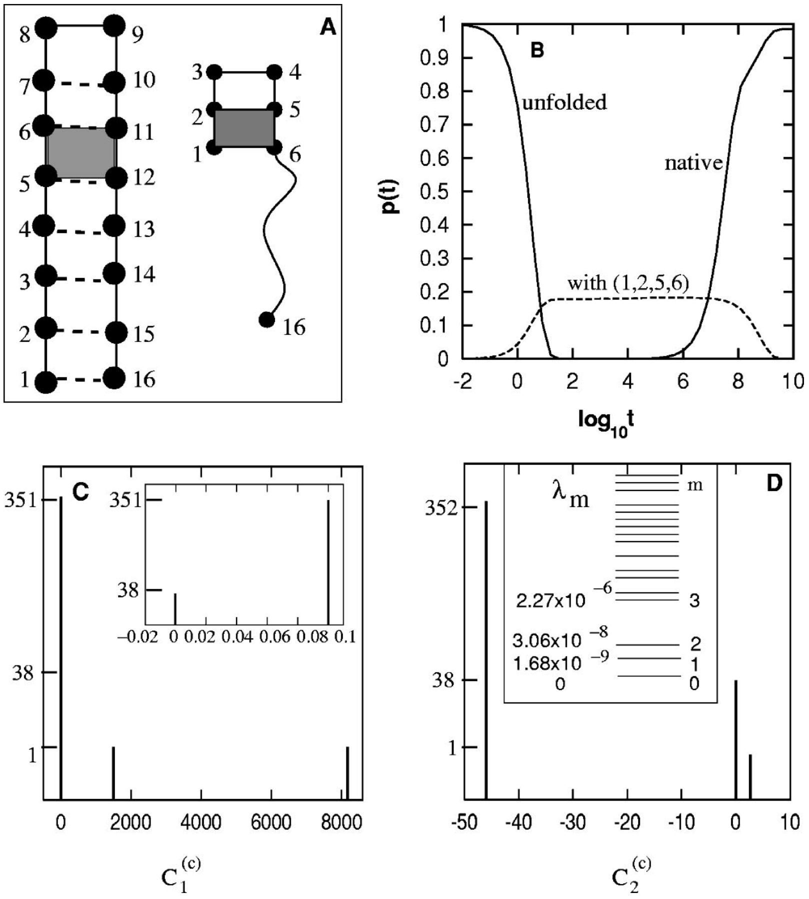

FIG. 7. (A) The native structure and the (designed) trapped structures with the shaded (1,2,5,6) stack. The formation of the shaded (5,6,11,12) stack is an (designed) on-pathway rate-limiting folding step. (B) The populational kinetics in the folding process. The kinetic intermediates correspond to the transient accumulation of conformations with (1,2,5,6). The native population shows a biphasic kinetics. (C) The histogram of $C_1^{(c)}$ for the slow mode $m_s=1$. (D) The eigenvalue spectrum (inset) and the histogram of $C_1^{(c)}$ for the slow mode $m_s=2$.

(b) $t\sim\lambda_1^{-1}\sim\lambda_2^{-1}\sim10^9$: the molecule misfold and refold to the native state by breaking the mis-formed non-native base stacks through the slow modes ($m_s=1,2$).

E. Mixture of on- and off-pathway rate-limiting steps

Model. Our approach can be applied to complex folding process involving both on-pathway and off-pathways rate-limiting steps. We assume $(\Delta S),\Delta H)=(-3,-4)$ and $(-10,-4)$ for the formation of (1,2,5,6) and (5,6,11,12), respectively [see Fig. 7)(A)]. Therefore, the disruption of the non-native base stack (1,2,5,6) and the formation of native base stack (5,6,11,12) are expected to be an off-pathway and an on-pathway rate-limiting step, due to the large entropic barrier and the large enthalpic barrier, respectively.

Eigenvalue spectrum. At temperature $T=0.2<T_m$, the native structure is hairpin-like. The first five nonzero eigenvalues of the rate matrix are $-1.68\times10^{-9},-3.06\times10^{-8}$, $-2.27\times10^{-6},-4.89\times10^{-6},-5.17\times10^{-6}$. There is clearly a gap between the second and third nonzero eigenvalue, corresponding to two slow modes $m_s=1,2$ [see Fig. 7(D)].

$|C_{m_s}^{(c)}|$**–distribution.** For the $m_s=1$ slow mode, the weighting coefficient $C_1^{(c)}$ values for different initial folding conformation c's, as shown in Fig. 7(C), are clustered around four discrete values $-0.000\,13$, 8161, 1496, and 0.1 for c = conformations with the (5,6,11,12), the conformation with (1,2,5,6), the fully unfolded state, and all other conformations, respectively. Since the fully unfolded state U does not belong to the group of the largest weighting coefficient, there must exist an off-pathway rate-limiting step in the folding process. The conformations of the largest $|C_1^{(c)}|$ value (=8161) contain a single base stack (1,2,5,6). Therefore, the breaking of the (1,2,5,6) stack is an off-pathway rate-limiting step.

For the $m_s=2$ slow mode, the $C_2^{(c)}$ coefficients also show a discrete three-value distribution around 0.021, 2.7, and -46 [see Fig. 7(D)] for initial conformations c = conformations with (5,6,11,12), the conformation with (1,2,5,6), and all other conformations (including the unfolded state), respectively. Because the fully unfolded state belongs to the group of conformations of the largest $|C_1^{(c)}|$ value (=46), there must be an on-pathway rate-limiting step. All the conformations of the smallest $|C_1^{(c)}|$ value (=0.021) contain a common base stack (5,6,11,12) Therefore, the formation of (5,6,11,12) is an on-pathway rate-limiting step.

Kinetic partitioning. For the folding from the fully unfolded state U, the likelihood for the molecules be trapped in the $r=(1,2,5,6)$ kinetic trap is $C_{m_s}^{(U)}n_{m_s}^{(r)}\simeq18\%$ ($m_s=1$), and the likelihood for the molecules to fold through the formation of the (5,6,11,12) stack is 75% ($m_s=2$). Therefore, most of the molecules will avoid the (1,2,5,6) trap and fold quickly after the (5,6,11,12) stack is formed, and the rest of the molecules will be trapped in the state with the (1,2,5,6) stack and need to break the (1,2,5,6) base stack before refolding by passing through the kinetic barrier for the formation of the (11,12,15,16) base stack. Such a "kinetic partitioning"[48] process causes a two-time scale folding kinetics shown in the populational kinetic curves in Fig. 7(B).

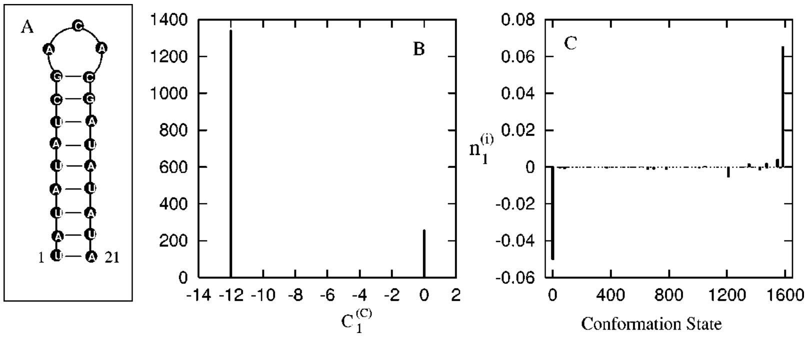

FIG. 8. (A) The native structure and 21-nt RNA hairpin-forming sequence (see Ref. 40). (B) The histogram of $C_1^{(c)}$ for the slow mode $m_s=1$ at $T=40\ °C$. The y axis represents the number of chain conformation c's that have the $C_1^{(c)}$ value represented by the x axis. (C) The vector components of the eigenvector of the slowest mode $m_s=1$ (see Ref. 40). The x-axis represents the conformations, and the y axis represents the corresponding components in the eigenvector.

V. APPLICATIONS TO RNA HAIRPIN FOLDING KINETICS

In the above sections, we have applied the method to the folding kinetics of model hairpin-like conformations. We have chosen such a simplified model in order to have an unambiguous way to test the theory, and to demonstrate the connections between the master equation and the complex on- and off-pathway kinetics and the kinetic partitioning mechanism. In this section, we apply the method to predict RNA hairpin folding kinetics using realistic chain models and experimentally measured energy and entropy parameters.[40,43]

The sequence and the native structure of the molecule are shown in Fig. 8(A).[40] This 21-nt RNA sequence has totally 1603 possible native and non-native secondary structures. The temperature dependence of the heat capacity shows a melting temperature of $T_m=62\ °\mathrm{C}$.[40] In the following, we investigate the rate-determining steps for the folding kinetics at folding condition $T=40\ °\mathrm{C}<T_m$ and at unfolding condition $T=90\ °\mathrm{C}>T_m$, respectively.

T=40 °C. The native state occupies 86.3% fraction population in thermal equilibrium at $T=40\ °\mathrm{C}$. The eigenvalue spectrum shows that it has a single slow mode $m_s=1$. As shown in Fig. 8(B), $C_{m_s}^{(c)}$ for the slow mode $m_s=1$ shows a 2-value distribution: $C_1^{(c)}\sim 0.007$ for all the conformations (c) that have the UCGA base stack formed, and $C_1^{(c)}\sim -12.6$ for all the conformations without the UCGA stack. Therefore, the formation of the base stack UCGA is the (on-pathway) rate-limiting folding step. Figure 8(C) shows the distribution of eigenvector components for the slow mode $m_s=1$. Only the denatured state [labeled as state 1 in Fig. 8(C)] and native state [labeled as state 1585 in Fig. 8(C)] are significantly populated, indicating that the dominant folding mode is the depletion of the fully extended conformations into the fully folded native conformation. But the eigenvector components do not directly reveal the rate-limiting steps. It is the $C_{m_s}^{(c)}$ distribution that unambiguously gives the rate-limiting steps. Physically, the formation of the UCGA stack involves the largest entropic loss, and thus is rate determining.

T=90 °C. The completely unfolded state occupies 98% of the total population in thermal equilibrium at $T=90\ °\mathrm{C}$. The eigenvalue spectrum shows that it has a single slow mode $m_s=1$. As shown in Fig. 9(A), $C_{m_s}^{(c)}$ for the slow mode $m_s=1$ shows a 2-value distribution: $C_1^{(c)}\sim -17.2$ for all conformations (c) that have the UCGA base stack formed, and $C_1^{(c)}\sim 0.05$ for all the conformations without the UCGA stack. Therefore, the breaking of base stack UCGA is the (on-pathway) rate-limiting step for the unfolding. Figure 9(B) shows the eigenvector distribution of the slowest mode $m_s=1$. There are multiple significantly populated states in the eigenvector spectrum. All these populated states have the UCGA stack. These (pre-equilibrated[40]) states reside inside a free energy well separated from the unfolded state by a high barrier arising from the large enthalpic cost to break the UCGA stack.[40]

The above results obtained from the current simple analytical method, for both the folding and unfolding kinetics, agree well with the conclusions derived from the previous computationally intensive brute force analysis for the folding pathways and energy landscapes[40] for the complete conformational ensemble.

VI. SUMMARY

The present transition state searching method can also be applied to the unfolding kinetics, where the rate-limiting steps involve the breaking of certain critical contacts. The rate-limiting steps in unfolding can be identified as the disruption of the common structure shared by the conformations (c) with the large value of $|C_m^{(c)}|$ for the slow modes (m_s).

A basic premise of our transition state searching method is the discrete distribution of the $C_{m_s}^{(c)}$ values for different folding conformations (c) for the slow modes (m_s). The separation in the $C_m^{(c)}$ values and the gap in the eigenvalue spectrum are related to each other. For a bumpy free energy landscape at low temperatures, there would be no dominant rate-limiting processes, and there would be no well defined

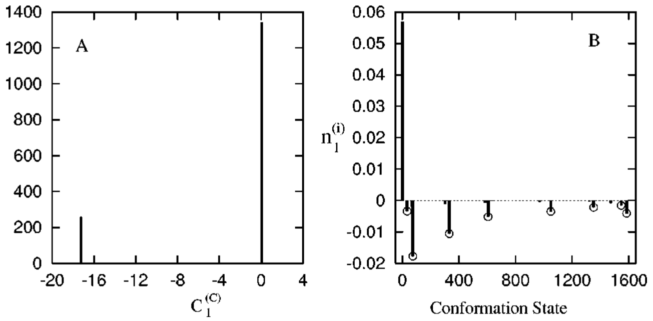

FIG. 9. (A) The histogram of $C_1^{(c)}$ for the slow mode $m_s=1$ at $T=90\ °C$. (B) The vector components of the eigenvector of the slowest mode $m_s=1$ (see Ref. 40).

slow and fast modes, and, in general, there would no separation in the $C_{m_s}^{(c)}$ value for different conformations for the slow modes.

Great advances have been made on the master equation approach to the folding kinetics.[9,31,40] The analytical method introduced in this work does not involve Monte Carlo simulations, and can identify the rate-limiting steps in an unambiguous and rigorous way. We have used a realistic RNA hairpin folding problem as a paradigm, but the rate-limiting searching method is general and is applicable to more complex RNA folding problems. Using efficient computational algorithms (e.g., to cluster conformations into pre-equilibrated macrostates,[30] the Lanczos algorithm,[49] and the truncation technique for diagonalization of large matrices[31]), we may extend the method to treat large RNAs and other biopolymer molecules.

ACKNOWLEDGMENTS

This work has been supported by grants from the American Heart Association (National Center) and the Petroleum Research Fund.

[1] C. J. Cerjan and W. H. Miller, J. Chem. Phys. **75**, 2800 (1981).
[2] J. Baker, J. Comput. Chem. **7**, 385 (1986).
[3] H. B. Schlemiels, J. Comput. Chem. **3**, 214 (1982).
[4] F. Jensen, J. Chem. Phys. **102**, 6706 (1995).
[5] D. K. Klimov and D. Thirumalai, Proteins **43**, 465 (2001).
[6] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).
[7] R. Du, V. S. Pande, A. Y. Rosenberg, T. Tanaka, and E. I. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).
[8] V. S. Pande and D. S. Rokhsar, Proc. Natl. Acad. Sci. U.S.A. **96**, 1273 (1999).
[9] S. B. Ozkan, I. Bahar, and K. A. Dill, Nat. Struct. Biol. **8**, 765 (2001).
[10] K. A. Dill and H. S. Chan, Nat. Struct. Biol. **4**, 10 (1997).
[11] H. Isambert and E. D. Siggia, Proc. Natl. Acad. Sci. U.S.A. **97**, 6515 (2000).
[12] A. Ansari, S. V. Kuznetsov, and Y. Q. Shen, Proc. Natl. Acad. Sci. U.S.A. **98**, 7771 (2001).
[13] V. Munoz, P. A. Thompson, J. Hofrichter, and W. A. Eaton, Nature (London) **390**, 196 (1997).
[14] K. M. Westerberg and C. A. Floudas, J. Chem. Phys. **110**, 9259 (1999).
[15] J. N. Onuchic, N. D. Socci, Z. LutheySchulten, and P. G. Wolynes, Fold Des **1**, 441 (1996).
[16] Z. Y. Guo and D. Thirumalai, Fold Des **2**, 377 (1997).
[17] P. Wolynes, Fold Des **3**, R107 (1998).
[18] Z. Y. Guo and D. Thirumalai, Biopolymers **36**, 83 (1995).
[19] D. Thirumalai and Z. Y. Guo, Biopolymers **35**, 137 (1995).
[20] D. K. Klimov and D. Thirumalai, J. Mol. Biol. **282**, 471 (1998).
[21] N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich, J. Mol. Biol. **296**, 1183 (2000).
[22] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, Biochemistry **33**, 10026 (1994).
[23] H. Nymeyer, N. D. Socci, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **97**, 634 (2000).
[24] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **103**, 4732 (1995).
[25] H. S. Chan and K. A. Dill, Proteins: Struct., Funct., Genet. **30**, 2 (1998).
[26] K. A. Dill, Protein Sci. **8**, 1166 (1999).
[27] H. S. Chan and K. A. Dill, J. Chem. Phys. **100**, 9238 (1994).
[28] C. Guo, H. Levine, and D. A. Kessler, Proc. Natl. Acad. Sci. U.S.A. **97**, 10775 (2000).
[29] R. Czerminski and R. Elber, J. Chem. Phys. **92**, 5580 (1990).
[30] O. M. Becker and M. Karplus, J. Chem. Phys. **106**, 1495 (1997).
[31] M. Cieplak, M. Henkel, J. Karbowski, and J. R. Banavar, Phys. Rev. Lett. **80**, 3654 (1998).
[32] R. Zwanzig, Proc. Natl. Acad. Sci. U.S.A. **94**, 148 (1997).
[33] E. I. Shakhnovich and A. M. Gutin, Europhys. Lett. **9**, 569 (1989).
[34] J. G. Saven, J. Wang, and P. G. Wolynes, J. Chem. Phys. **101**, 11037 (1994).
[35] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).
[36] J. D. Bryngelson and P. G. Wolynes, J. Phys. Chem. **93**, 6902 (1989).
[37] Y.-J. Ye, D. R. Ripoll, and H. A. Scheraga, Comput. Theor. Polym. Sci. **9**, 359 (1999).
[38] Y.-J. Ye and H. A. Scheraga, in *Slow Dynamics in Complex Systems*, AIP Conf. Proc. No. 469 (AIP, New York, 1999), p. 452.
[39] M.-H. Hao and H. A. Scheraga, J. Phys. Chem. **107**, 8089 (1997).
[40] W. B. Zhang and S. J. Chen, Proc. Natl. Acad. Sci. U.S.A. **99**, 1931 (2002).
[41] M. Tacker, W. Fontana, P. F. Stadler, and P. Schuster, Eur. Biophys. J. **23**, 29 (1994).
[42] C. K. Hall and E. Helfand, J. Chem. Phys. **77**, 3275 (1982).
[43] M. J. Serra and D. H. Turner, Methods Enzymol. **259**, 242 (1995).
[44] W. B. Zhang and S. J. Chen, J. Chem. Phys. **114**, 7669 (2001).
[45] E. Tostesen, S. J. Chen, and K. A. Dill, J. Phys. Chem. B **105**, 1618 (2001).
[46] S. J. Chen and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **97**, 646 (2000).
[47] S. J. Chen and K. A. Dill, J. Chem. Phys. **109**, 4602 (1998).
[48] D. Thirumalai and S. A. Woodson, Acc. Chem. Res. **29**, 433 (1996).
[49] J. K. Cullum and R. A. Willoughby, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations* (Birkhauser, Boston, 1985).