Analyzing the biopolymer folding rates and pathways using kinetic cluster method

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

(Received 20 March 2003; accepted 1 August 2003)

A kinetic cluster method enables us to analyze biopolymer folding kinetics with discrete rate-limiting steps by classifying biopolymer conformations into pre-equilibrated clusters. The overall folding kinetics is determined by the intercluster transitions. Due to the complex energy landscapes of biopolymers, the intercluster transitions have multiple pathways and can have kinetic intermediates (local free-energy minima) distributed on the intercluster pathways. We focus on the RNA secondary structure folding kinetics. The dominant folding pathways and the kinetic partitioning mechanism can be identified and quantified from the rate constants for different intercluster pathways. Moreover, the temperature dependence of the folding rate can be analyzed from the interplay between the stabilities of the on-pathway (nativelike) and off-pathway (misfolded) conformations and from the kinetic partitioning between different intercluster pathways. The predicted folding kinetics can be directly tested against experiments. © *2003 American Institute of Physics.* [DOI: 10.1063/1.1613255]

I. INTRODUCTION

The conformations and conformational transitions of biopolymers, such as RNAs, DNAs, and proteins, play critical roles in their biological functions. It is extremely important to develop a theory that can predict and analyze the kinetics of biopolymer conformational transitions. Unlike small molecules, biopolymers are mostly flexible chain molecules and have a large number of accessible chain conformations. The large number of chain conformations form an exceedingly complex energy landscape for a biopolymer. For example, a chain molecule can fold and unfold along many different routes that connect the initial and final states. Moreover, due to the complex interplay between entropy and enthalpy, the chain can either fold along smooth downhill fast pathways, or be trapped by kinetic intermediates (traps) that must be disrupted in the folding process.[1–6] Therefore, in order to treat realistic biopolymers, a folding kinetics theory must be able to account for the large number of chain conformations, the multiple kinetic pathways, and the kinetic traps.

Predicting the detailed folding kinetics from the first principle analytical calculations is generally not yet possible for realistic biopolymers. One of the key problems in developing a folding kinetics theory is how to treat the enormously large conformational ensemble. For example, we are not able to quantitatively analyze the RNA folding kinetics based on the complete ensemble of RNA secondary structures. In this paper, by dividing the full conformational ensemble into several interconnected clusters (the kinetic cluster method), we develop a rigorous statistical mechanical framework to predict the detailed RNA secondary structure folding kinetics, including the folding and unfolding rates, dominant pathways, possible kinetic intermediates, and the temperature dependence of the folding kinetics. With the kinetic cluster method, we can make reliable predictions for the folding kinetics through analytical calculations, and the predicted results can be directly applied to the experiments.

A number of studies have applied the master equation approach to describe the protein and RNA folding kinetics.[7–19] Consider a chain molecule that has Ω conformational states; let P_i be the fractional population (probability) of state i. The time evolution of P_i obeys the following master equation:

$$\frac{dP_i}{dt}=\sum_j (k_{j\to i}P_j - k_{i\to j}P_i),$$

for $i=1,2,\ldots,\Omega$, where $k_{i\to j}$ and $k_{j\to i}$ are the rate constants for transitions from state i to j and from j to i, respectively. The rate constants form the rate matrix. For a given initial condition of the populational distribution, the populational kinetics $\mathbf{P}(t)=\mathrm{col}\,(P_1(t),P_2(t),P_3(t),\ldots)$ is determined by the linear superposition of the eigenmodes of the rate matrix

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

where $-\lambda_m$ and $\mathbf{n}_m$ are the mth eigenvalue and eigenvector of the rate matrix, respectively, and C_m is the coefficient determined by the initial condition at time $t=0$. Each eigenmode $(-\lambda_m,\mathbf{n}_m)$ represents a kinetic mode of rate λ_m.

The master equation description of the folding kinetics is general. It considers the full ensemble of the chain conformational states and accounts for the transitions between each and every pair of the kinetically connected states. However, since the number of chain conformations Ω increases expo-

a)Author to whom correspondence should be addressed. Electronic mail: chenshi@missouri.edu

nentially as the chain length grows, the applicability of the master equation approach for realistic biopolymers is strongly limited by the large size ($\Omega \times \Omega$) of the rate matrix. But, if there exist discrete rate-limiting steps for the kinetic process, it would be possible to ''renormalize'' the conformational space into a number of clusters. The large ensemble of chain conformations can thus be drastically reduced into a much smaller number of conformational clusters.[20–22]

Different clusters are separated by the rate-limiting steps. If the rate-limiting steps involve sufficiently high kinetic barrier, the microstates within each cluster would have sufficient time to equilibrate and form a macrostate (in local equilibrium) before crossing the intercluster barriers to enter other (kinetically neighboring) clusters. The transitions between different clusters (macrostates) determine the overall folding kinetics of the molecule.

One might use the classical transition state theory to estimate the intercluster rate constants. The classical standard transition state theory, originally developed for chemical reaction dynamics, is based on two assumptions:[23] reactants and products are thermal equilibrium ensembles, and molecules that pass through the transition state will proceed to the products. The classical transition state theory is not directly applicable to the intercluster dynamics in biopolymer folding. First, there are multiple parallel pathways connecting different clusters and thus the intercluster kinetics is more convoluted. Second, often the pathways involve kinetic intermediates between the clusters. The kinetic intermediates (and traps) can cause the ''rebounds'' effect: the molecule may proceed to cross over an intercluster barrier but then immediately find itself trapped in a local minima, causing the molecule to recross the intercluster barrier.

A number of attempts have been made to model the intercluster kinetics.[20–22,24–26] In general, two types of methods have been proposed. In the first type of approach, the intercluster rate is computed based on the transition state with the lowest barrier. For each pair of initial and final states, the optimized pathway is used to estimate the rate constant. In the second type of approach, conformations interconvertible through barrierless transitions are classified as a cluster, and the rate constants are calculated based on all the possible intercluster kinetic pathways. Our present kinetic cluster method is based on the pre-equilibrated macrostates, which includes the barrierless conformational cluster used in the previous model[25,26] as a special subset. So, the present approach is more general. From the intercluster rate constants, we construct the reduced rate matrix. From the eigenvalues and eigenvectors of the reduced rate matrix, we can analyze the folding/unfolding rates and the pathways for the overall kinetics.

The folding kinetics analysis is often extremely convoluted due to the large number of possible conformations and pathways and the complex shape of the free-energy landscape. With the kinetic cluster approach developed here, we can find the dominant folding pathways. Also, from the interplay between the stabilities of the nativelike and the misfolded states and from the kinetic trapping effects, we can predict the temperature dependence of the folding rate from analytical calculations. Moreover, the present method may be a first step for the development of a method to treat biopolymer folding kinetics with long chain length. A master equation approach to the long chain folding kinetics has been limited by the large conformational ensemble, while the present kinetic cluster method can reduce the large conformational ensemble into several clusters.

The pre-equilibration and cluster formation have been observed in previous experiments and computer simulations,[27–38] and the pre-equilibrated macrostates have been used to quantitatively analyze the chemical kinetics and other small molecule systems.[33–38] The present kinetic cluster approach is a systematic and analytical treatment for complex biopolymer folding kinetics, which has strong sequence dependence, and involves multiple pathways, intermediates, traps, and networks of clusters. The present kinetic cluster approach would be a step forward toward a statistical mechanical framework for the *ab initio* predictions for the folding rates, pathways, traps (intermediates), and their temperature dependences from the sequence.

II. KINETIC CLUSTER ANALYSIS FOR BIOPOLYMER FOLDING KINETICS

A. Intercluster kinetics

We assume that according to the rate-limiting steps, the conformational ensemble can be divided into two clusters denoted by C and N, such that the intracluster transitions ($C \leftrightarrow C$ and $N \leftrightarrow N$) are much faster than the intercluster transitions ($C \leftrightarrow N$). We also assume that cluster C has Ω_C conformations: $C_1, C_2, \ldots, C_{\Omega_C}$, and cluster N has Ω_N conformations: $N_1, N_2, \ldots, N_{\Omega_N}$.

In general, there are multiple pathways that can connect the clusters. We use $\omega_{C \leftrightarrow N}$ to denote the number of the intercluster pathways, and use $C_i \leftrightarrow N_i$ to denote the ith pathway, where conformations C_i and N_i are connected by a kinetic move. We call such conformations C_i and N_i on the intercluster pathways ''pathway conformations,'' and call the other conformations ''nonpathway conformations.''

Considering that C_i and N_i occupy certain fractional populations, the effective contribution of the $C_i \leftrightarrow N_i$ pathway to the overall intercluster transition rates is given by the following equations:

$$k^{\text{eff}}_{C_i \to N_i} = P_{C_i} \; k_{C_i \to N_i}; k^{\text{eff}}_{N_i \to C_i} = P_{N_i} k_{N_i \to C_i}; \tag{2}$$

where P_{C_i} and P_{N_i} are the fractional populations of C_i and N_i in the respective clusters and are determined by the following equilibrium distribution:

$$P_{C_i} = e^{-(G_{C_i} - G_C)/k_B T}; \;\; P_{N_i} = e^{-(G_{N_i} - G_N)/k_B T}, \tag{3}$$

where G_C and G_N are the free energy of the conformational ensemble in clusters C and N, respectively,

$$G_C = -k_B T \ln\left(\sum_{j=1}^{\Omega_C} e^{-G_{C_j}/k_B T} \right);$$

$$G_N = -k_B T \ln\left(\sum_{j=1}^{\Omega_N} e^{-G_{N_j}/k_B T} \right). \tag{4}$$

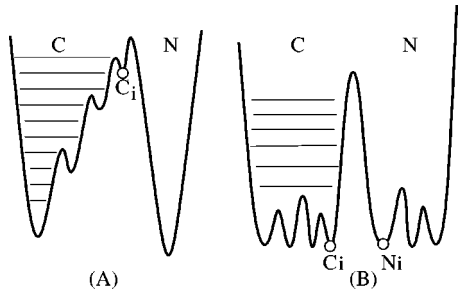

FIG. 1. Free-energy landscape and the formation of a cluster. (A) The pathway conformation C_i has high free energies and small populations and thus has a slow rate to escape from cluster C; (B) the intercluster pathways ($C_i \rightarrow N_i$) have high free-energy barriers and slow rates.

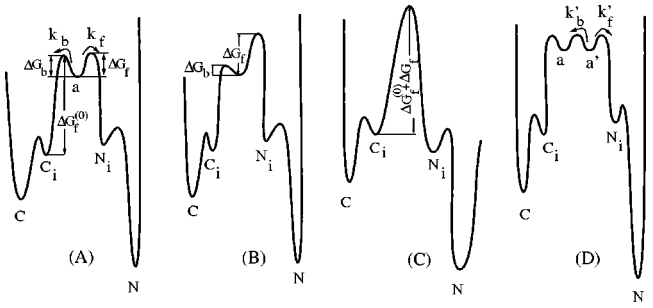

FIG. 2. An intercluster pathway ($C_i \leftrightarrow N_i$) that involves one local minimum a_i: (A) ΔG_b and ΔG_f are on the same order of magnitude; (B) $\Delta G_b \ll \Delta G_f$; (C) $\Delta G_b = 0$ (cooperative transition between C_i and N_i); and (D) two local minima a_i and a_i'.

The total intercluster transition rate can be computed as the sum over all the $\omega_{C\leftrightarrow N}$ pathways

$$k_{C\rightarrow N} = \sum_{i=1}^{\omega_{C\leftrightarrow N}} k^{\mathrm{eff}}_{C_i\rightarrow N_i}; \quad k_{N\rightarrow C} = \sum_{i=1}^{\omega_{C\leftrightarrow N}} k^{\mathrm{eff}}_{N_i\rightarrow C_i}. \tag{5}$$

B. Formation of the kinetic clusters

In general, the clusters should satisfy the condition that the intracluster transitions are much faster than the intercluster transitions, such that each cluster is a quasiclosed system. From Eqs. (2)–(5), a slow intercluster rate $k_{C\rightarrow N}$ and $k_{N\rightarrow C}$ can arise from (a) small number $\omega_{C\leftrightarrow N}$ of the intercluster pathways; (b) small fractional populations P_{C_i} and P_{N_i} for the pathway conformations, i.e., high free energies of C_i (and N_i) in the respective clusters [see Fig. 1(A)]; (c) small rate constants $k_{C_i\rightarrow N_i}$ and $k_{N_i\rightarrow C_i}$ for each intercluster pathway, i.e., high kinetic barrier between C_i and N_i [see Fig. 1(B)].

The pathway conformations effectively form the boundary of a cluster in the conformational space. In order for a cluster to form a pre-equilibrated macrostate, the rates for the (pathway) conformations to escape from the cluster must be small as compared to the rates to enter the cluster. Mathematically, this condition can be expressed by the following inequalities for the rate constants related to the pathway conformations C_i and N_i ($i=1,2,\ldots,\Omega_{C\leftrightarrow N}$):

$$\sum_{N_j\in N} k_{C_i\rightarrow N_j} \ll \sum_{C_j\in C} k_{C_i\rightarrow C_j}; \quad \sum_{C_j\in C} k_{N_i\rightarrow C_j} \ll \sum_{N_j\in N} k_{N_i\rightarrow N_j}. \tag{6}$$

The clusters in Figs. 1(A) and 1(B) represent two different scenarios that satisfy the above conditions.

The classification of the conformations into clusters depends on the temperature, solvent condition, etc., because different conditions may have different rate-limiting steps that divide the clusters. For example, under a strong folding condition, the overall relaxation kinetics would be dominated by the folding process. In this case, the rate-limiting steps can be (a) the slow steps in the formation of native intrachain contacts and (b) the slow steps in the disruption of nonnative contacts. Under conditions that the unfolding process becomes dominant, the rate-limiting steps can be the slow disruption of the native contacts.

C. Dominant intercluster pathways

Suppose the folded native state is in cluster N. The probability for a molecule to fold through a pathway $C_i \rightarrow N_i$ is determined by the ratio $k^{\mathrm{eff}}_{C_i\rightarrow N_i}/k_{C\rightarrow N}$, where $k^{\mathrm{eff}}_{C_i\rightarrow N_i}$ is the contribution from the $C_i \rightarrow N_i$ [see Eq. (2)] and $k_{C\rightarrow N}$ is the total rate for all the pathways [see Eq. (5)], and the probability for an unfolding reaction through pathway $N_i \rightarrow C_i$ is determined by the ratio $k^{\mathrm{eff}}_{N_i\rightarrow C_i}/k_{N\rightarrow C}$. The dominant pathways in the overall kinetics are the most probable $C_i \rightarrow N_i$ and $N_i \rightarrow C_i$ transitions. Since $k^{\mathrm{eff}}_{C_i\rightarrow N_i}$, $k^{\mathrm{eff}}_{N_i\rightarrow C_i}$, and $k_{C\rightarrow N}$, $k_{N\rightarrow C}$ are strongly temperature dependent [see Eqs. (2)–(5)], the dominant pathways can be quite different for different temperatures.

D. Effect of the kinetic intermediates on the intercluster kinetics

Due to the complex free-energy landscape of biopolymers, there may exist kinetic intermediates distributed on the intercluster pathways. These kinetic intermediate play an important role in determining the rate and the dominant pathways for the intercluster transitions and for the overall kinetics of the system.

In Fig. 2 we show a schematic one-dimensional free-energy landscape that involves one kinetic intermediate [Figs. 2(A)–2(C)] and multiple kinetic intermediates [Fig. 2(D)] on an intercluster pathway. Physically, if cluster C represents a macrostate for non-native conformations that form a deep kinetic trap, Fig. 2(A) shows that after the molecule is detrapped from C through conformation C_i, it immediately finds itself trapped in a local minimum a before proceeding to fold into the native cluster N (through, e.g., state N_i).

The free-energy landscape and the rate constants related to a determine that a cannot be treated as a pathway conformation in either cluster C or cluster N [see Eq. (6)]. One can treat the kinetic intermediate as a separate state, and solve the master equation for the multicluster system including local minima a's as separate states. However, given the large number of intercluster pathways, such an approach could be computationally difficult and the analysis for the results would be quite convoluted in general. Therefore, it is useful

to derive an analytical expression for the folding rates based on simple physical analysis.

1. The folding rate

We use Fig. 2(A) to illustrate the general method. Suppose the native state is in cluster N; then, the transition $C \rightarrow N$ represents a folding process, and the folding rate k_F is equal to the intercluster transition rate for $C \rightarrow N$. As shown in Fig. 2(A), since a certain fraction of population that enters the intermediate state a from state C_i would quickly rebound to re-enter cluster C, the effective populations for the forward folding process would be reduced.

We assume that the intermediate state a is not sufficiently stable to form a deep kinetic trap, and there is thus no significant accumulation in the population for state a (steady-state approximation). In this case, the fractional population that flows back to cluster C from the state a is determined by the rate constants k_b and k_f [see Fig. 2(A)] for the backward rebound transitions and the forward folding transitions, respectively: $k_\mathrm{b}/(k_\mathrm{f}+k_\mathrm{b})$, and the fractional population that flows into the native cluster N is $k_\mathrm{f}/(k_\mathrm{f}+k_\mathrm{b})$. These factors are similar to the transmission coefficients in the transition state theory for chemical kinetics.[39] Considering the multiple pathways in the rebound and folding transitions, k_b and k_f can be the total rate for all the possible backward rebound transitions (from a) and for all the possible forward folding transitions (from a), respectively: $k_\mathrm{b}=\Sigma_{C_j \in C} k_{a \rightarrow C_j}$; $k_\mathrm{f}=\Sigma_{N_j \in N} k_{a \rightarrow N_j}$.

Considering the distribution of the fractional population P_{C_i} of state C_i in cluster C [see Eq. (3)], we obtain the following effective rate constant $k_{C_i \rightarrow N}$ for the folding transition $C_i \rightarrow N$ through state a:

$$k_{C_i \rightarrow N}^{(\mathrm{eff})} = P_{C_i} k_{C_i \rightarrow a} k_\mathrm{f}/(k_\mathrm{f}+k_\mathrm{b}). \tag{7}$$

The overall folding rate k_F is given by the sum over all the possible pathways between C and N:

$$k_F = \sum_{i=1}^{\omega_{C \leftrightarrow N}} k_{C_i \rightarrow N}^{(\mathrm{eff})} = \sum_{i=1}^{\omega_{C \leftrightarrow N}} P_{C_i} k_{C_i \rightarrow a} k_\mathrm{f}/(k_\mathrm{f}+k_\mathrm{b}). \tag{8}$$

Similarly, the unfolding rate k_U for the $N \rightarrow C$ transition is given by

$$k_U = \sum_{i=1}^{\omega_{C \leftrightarrow N}} P_{N_i} k_{N_i \rightarrow a} k_\mathrm{b}/(k_\mathrm{f}+k_\mathrm{b}). \tag{9}$$

As we will show later in the following sections, Eq. (8) and Eq. (9) give quite accurate estimations for the folding and unfolding rates for the intermediate states-mediated intercluster transitions.

The above analysis for the rebounds effect on the intercluster kinetics and the rates k_F and k_U is based on the assumption that the local minimum (a) is not a deep trap and thus there is no significant trapping of the population. Equation (8) and Eq. (9) show that k_F and k_U depend on the intermediates a through the relative ratio $k_\mathrm{f}/k_\mathrm{b}$ rather than the absolute values of k_f and k_b. In terms of the free energies shown in Figs. 2(A)–2(C), k_F and k_U depends only on the difference in the free-energy barriers, $\Delta G_f - \Delta G_b$, not the absolute values of ΔG_f (for the forward transition $a \rightarrow N_i$) or ΔG_b (for the backward transition $a \rightarrow C_i$). However, if the barriers ΔG_f and ΔG_b are very high, a would become very stable and the trapping effect would be significant. In the deep trapping case, the folding and unfolding rate k_F and k_U would depend on the absolute values of ΔG_f and ΔG_b. In fact, as the local minimum a becomes deeper, due to the populational accumulation and the related kinetic pause, the folding and unfolding would be slower.

2. Folding rate and the stability of the intermediate states

For a schematic one-dimensional free-energy landscape shown in Fig. 2(A), we assume that the rate constants are determined by the corresponding kinetic barriers as the following:

$$k_{C_i \rightarrow a} = e^{-\Delta G_f^{(0)}/k_B T}; \quad k_\mathrm{f} = e^{-\Delta G_f/k_B T}; \quad k_\mathrm{b} = e^{-\Delta G_b/k_B T}.$$

The rate for the intermediate state-mediated intercluster transitions, according to Eq. (8), is given by

$$k_F = P_{C_i} k_{C_i \rightarrow a} \frac{k_\mathrm{f}}{k_\mathrm{f}+k_\mathrm{b}} = P_{C_i} \frac{k_{C_i \rightarrow a}\ k_\mathrm{f}}{e^{-\Delta G_f/k_B T} + e^{-\Delta G_b/k_B T}}. \tag{10}$$

For a fixed (forward) kinetic barrier ΔG_f, a higher barrier ΔG_b for the recrossing transition leads to a reduced rate for the rebound process, which effectively causes a more productive forward folding process. For a sufficiently low barrier $\Delta G_b \ll \Delta G_f$, as shown in Fig. 2(B), the intermediate state a can quickly exchange populations with conformations (C_i) in cluster C to reach local equilibrium. As a result, a can be included in cluster C as a pathway conformation. In the limiting case with $\Delta G_b \rightarrow 0$, as shown in Fig. 2(C), the intermediate state a does not exist, and the intercluster transition becomes a cooperative process with no intermediates. The folding rate for the cooperative process is $k_F^{\mathrm{coop}} = e^{-(\Delta G_f^{(0)} + \Delta G_f)/k_B T} = P_{C_i}\ k_{C_i \rightarrow a}\ k_F$.

From the above expressions for k_F and k_F^{coop}, we find that, for processes with the same total folding barrier $\Delta G_f^{(0)} + \Delta G_f$, the noncooperative process (without a) could have a faster folding rate than the cooperative process ($k_F > k_F^{\mathrm{coop}}$) provided that $e^{-\Delta G_f/k_B T} + e^{-\Delta G_b/k_B T} < 1$. Physically, the above condition requires sufficiently high free-energy barriers ΔG_f and ΔG_b, i.e., sufficiently stable intermediate (but not stable enough to become a deep kinetic trap). The physical mechanism of such a local minima-mediated folding "acceleration" is due to the slowdown of the rebound process due to the kinetic barrier ΔG_b for the recrossing, as compared with the cooperative process with no such kinetic barrier ($\Delta G_b = 0$) for the recrossing process.

The accelerated folding rate due to the entropic effect of the intermediates in the transition states was previously reported.[40] It was found that for different landscapes with the same uphill barrier barrier $\Delta G_f^{(0)}$ in Fig. 2, stabilization of the intermediates could lead to a faster folding rate. The folding acceleration reported in that work is different from that reported here: First, we compare different folding processes with the same total folding barrier $\Delta G_f^{(0)} + \Delta G_f$ rather

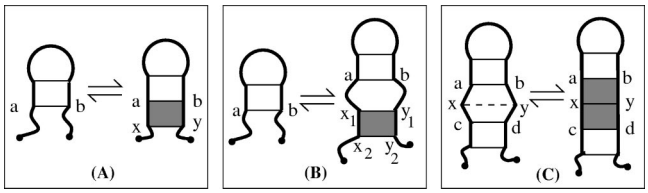

FIG. 3. A kinetic move is defined as the formation/disruption of a stack (two adjacent intrachain contacts): (x,a,b,y) in (A) and (x2,x1,y1,y2) in (B), or a stacked contact: (x,y) in (C).

than those with the same $\Delta G_f^{(0)}$. Second, the acceleration of folding here is due to the reduced rebounds and the enhanced productive folding rather than the enhanced entropic effects in the transition state ensemble.

We can generalize the above approach to treat processes involving multiple intermediates between clusters C and N. For two sequentially distributed intermediates a and a' shown in Fig. 2(D), we consider the following iterative process: a fraction of $k_b'/(k_f'+k_b')$ of the population in state a' would rebound to state a, and a fraction of $k_f/(k_f+k_b)$ of such rebound population would refold and re-enter state a'. Therefore, similar to Eq. (7), the effective rate constant $k_{C_i\to N}$ for the folding transition $C_i\to N$ is given by

$$k_{C_i\to N}^{(\mathrm{eff})}=P_{C_i}k_{C_i\to a}\left(\frac{k_f}{k_f+k_b}\right)\left(\frac{k_f'}{k_f'+k_b'}\right)\times\sum_{n=0}^{\infty}\left(\frac{k_f}{k_f+k_b}\frac{k_b'}{k_f'+k_b'}\right)^n,$$

and the total folding rate is given by Eq. (8) with the rate constant $k_{C_i\to N}^{(\mathrm{eff})}$ given above. After simplification, we find $k_{C_i\to N}^{(\mathrm{eff})}=P_{C_i}k_{C_i\to a}k_fk_b'/[(k_f+k_b)k_f'+k_bk_b']$, from which we obtain the following condition for the folding acceleration (i.e., $k_{C_i\to N}^{(\mathrm{eff})}<$ the cooperative folding rate $P_{C_i}k_{C_i\to a}\,k_f\,k_b'$): $(k_f+k_b)k_f'+k_bk_b'<1$.

III. ILLUSTRATIVE CALCULATIONS FOR BIOPOLYMER FOLDING PROBLEM

A. The model

Our purpose here is twofold: (i) to demonstrate the implementations of the kinetic cluster method for biopolymer folding and (ii) to validate the method through comparisons with the results from the original exact master equations with the complete conformational ensemble. We choose a simplified RNA hairpinlike folding model which allows the formation and disruption of all the possible hairpin conformations. We assume that the conformations are stabilized by the (sequence-dependent) stacking interactions, therefore, we can use stacks (=two adjacent intrachain contacts; see Fig. 3) to describe the conformational states. Unstacked intrachain contacts are unstable and can be disrupted quickly. As a result, two conformations that differ by unstacked intrachain contacts are classified as the same conformational state. To further simplify the calculation, we neglect the loop entropies. The total free energy of a conformation i, according to the nearest model,[41,42] is equal to the additive sum of the free energies for all the constituent stacks: $\Delta G_i=-\Sigma_{\text{all stacks}}(\Delta H-T\Delta S)$, where ΔS and ΔH are the entropic and enthalpic change upon the formation of the corresponding stack, respectively. The native state for a given sequence is the state with the lowest free energy.

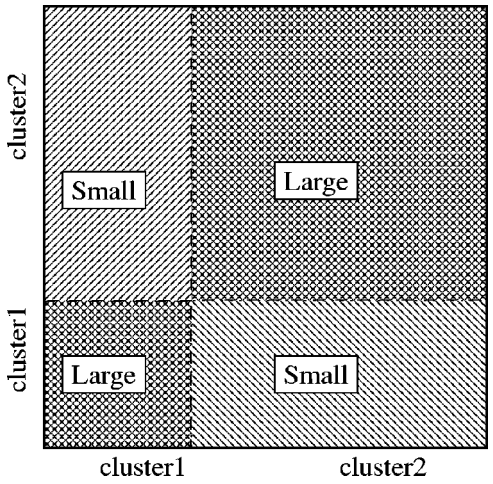

FIG. 4. The block-form rate matrix for a two-cluster system, where the intracluster transition rates are large and the intercluster transition rates are small (and sparsely distributed). The column and row indexes represent the conformational states.

The methodology developed here does not rely on any particular choice of kinetic moves. However, to be specific in the illustrative computation, we define a kinetic move to be the formation or breaking of a stack or a stacked intrachain contact; see Fig. 3. As shown in the figure, a kinetic move corresponds to the addition/deletion of either one stack [see Figs. 3(A) and 3(B)] or possibly two consecutive stacks [see Fig. 3(C)].

We assume that rate constant $k_{i\to j}$ for a kinetic move from state i to state j is determined by the free-energy barrier $\Delta G_{i\to j}^{\ddagger}$ for the kinetic move: $k_{i\to j}=e^{-\Delta G_{i\to j}^{\ddagger}/k_BT}$. Furthermore, we assume that the formation of a stack is rate limited by the associated entropic decrease $-\Delta S$ and the disruption of a stack is rate limited by the associated enthalpic increase ΔH. Therefore, $\Delta G^{\ddagger}$ is equal to $T\Delta S$ and ΔH for the formation and disruption of a stack, respectively, and the rate constants are given by

$$k_{i\to j}=(e^{-\Delta S},e^{-\Delta H/T})$$

for the (formation, disruption) of the stack, (11)

where we have set $k_B=1$ to simplify the notation.

If at temperature T, the formation/disruption of a stack, e.g., (x,a,b,y) in Fig. 3, is distinctively slow as compared with the formation/disruption of all other stacks, the formation/disruption of stack (x,a,b,y) could be a rate-limiting step. As a result, conformations with and without this rate-limiting stack can be classified into two clusters, as illustrated in Fig. 1(B). Mathematically, for such well-separated distribution of the rate constants, the rate matrix can be transformed into block form[21] such that the conformations within each cluster would have sufficient time to equilibrate before the rate-limiting stack is formed or disrupted; see Fig. 4.

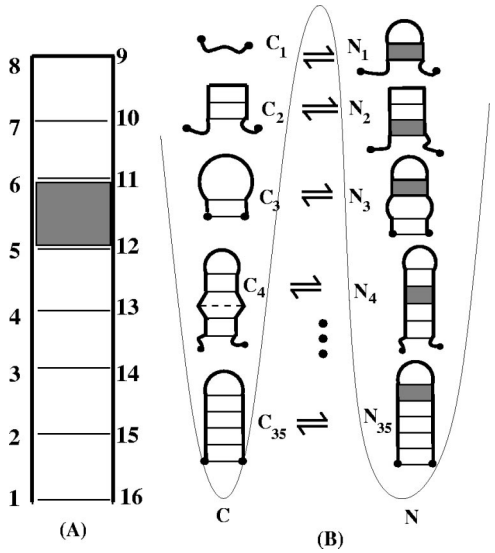

FIG. 5. (A) The native state and (B) the 35 intercluster pathways. Clusters N and C are for the conformations with and without the rate-limiting stack [shaded in (A)].

We allow the formation of all possible stacks, native or non-native, in the folding/unfolding processes. A stack is called native if the particular stack exists in the native state, and is called non-native otherwise. For the folding kinetics (a) the formation of a native stack with very large entropic reduction $-\Delta S$ is slow [see Eq. (11)] and is an on-pathway rate-limiting step; (b) the disruption of a non-native stack with very high enthalpic barrier ΔH is slow [see Eq. (11)] and is an off-pathway rate-limiting step. In general, for a given RNA sequence, we can identify the folding rate-limiting steps by examining the ΔS and ΔH parameters for all the possible stacks and other structural units. The formation of native structure with large ΔS values is on-pathway rate-limiting steps, and the non-native structures with large ΔH are off-pathway rate-limiting kinetic traps.

We choose a 16-nt RNA hairpin-forming chain, and assume a uniform entropic and enthalpic changes for the formation of all stacks: $(-\Delta S_0, -\Delta H_0) = (-5, -2)$, except for those special rate-determining stacks with distinctively larger enthalpic or entropic parameters. We denote the entropy and enthalpy parameters for the special rate-limiting stacks as ΔH^* and ΔS^*. For RNA secondary structures under 1M NaCl salt condition, the sequence-dependent enthalpic and entropic parameters for different stacks have been measured experimentally,[41] but here we use the simplified values for the purpose of model illustration, though the method developed here can be directly applied to realistic RNA molecules by using the realistic enthalpy and entropy parameters. For a 16-nt hairpin chain here, there are totally 391 stack-based hairpin conformational states. For the enthalpic and entropic parameters that we use, the native structure is a hairpin structure with a helical stem of seven consecutive stacks [see Fig. 5(A)].

We will apply the kinetic cluster method to thoroughly investigate the following representative folding scenarios for the simplified 16-nt chain model: folding with one on-pathway rate-limiting step, folding with one off-pathway kinetic trap and folding with two different on-pathway rate-limiting steps. The purpose of the third model calculation is to illustrate the application of the kinetic cluster analysis to a multiple rate-limiting steps folding process. Our strategy is to design different folding pathways by assigning different energy and entropy parameters (ΔH^* and ΔS^*) for the designed rate-determining stacks. For realistic RNA molecules, such an approach can effectively corresponds to the design of different sequences. For each designed model sequence, the kinetic cluster results will be tested against the exact solutions from the exact master equations for the complete 391 chain conformations.

B. Folding and unfolding kinetics with an on-pathway rate-limiting step

As shown in Fig. 5(A), we assume that the formation of the stack (5,6,11,12) involves a large entropic decrease $\Delta S^* (> \Delta S)$: $(-\Delta S^*, -\Delta H^*) = (-15, -6)$. For the given entropy and enthalpy parameters, the heat capacity melting curve shows[43–46] a single melting transition at the melting temperature $T_m = 0.4$. We first investigate the kinetics at temperature $T = 0.2$ $(< T_m)$. The native state has a fractional population of 98% at $T = 0.2$. Therefore, the relaxation process is predominantly a folding process.

1. Kinetic clusters

Because the ΔS^* and ΔH^* are larger than ΔS_0 and ΔH_0, respectively, according to Eq. (11), the formation/disruption of the native stack (5,6,11,12) is distinctively slower than the formation/disruption of other stacks. As shown in Fig. 6, the rate matrix can be transformed into a block form, and the conformations can be classified as two clusters: N and C for conformations with and without the rate-limiting stack (5,6,11,12). Clusters C and N have $\Omega_C = 353$ and $\Omega_N = 38$ conformations, respectively.

2. Intercluster kinetics

There are $\omega_{C \leftrightarrow N} = 35$ intercluster pathways, each corresponding to the addition or deletion of stack (5,6,11,12). Most pathways have rate constants $(e^{-\Delta S^*}, e^{-\Delta H^*/T}) = (e^{-15}, e^{-30})$ for the (formation, disruption) of the (5,6,11,12) stack; other pathways involve the simultaneous formation/disruption of two stacks [e.g., $C_4 \rightarrow N_4$ in Fig. 5(B); see also Fig. 3(C)] and thus have much slower rates $(e^{-(\Delta S_0 + \Delta S^*)}, e^{-(\Delta H_0 + \Delta H^*)/T}) = (e^{-17}, e^{-42.5})$. Equations (2)–(5) give the intercluster rate constants $k_{C \rightarrow N} = 2.088 \times 10^{-10}$ and $k_{N \rightarrow C} = 6.3 \times 10^{-16}$. Under the folding condition $T = 0.2$, the folding rate $k_F = k_{C \rightarrow N}$ is much larger than the unfolding rate $k_U = k_{N \rightarrow C}$ and the relaxation rate k_R is determined by $k_R = k_{C \rightarrow N} + k_{N \rightarrow C} \simeq 2.088 \times 10^{-10}$.

The above kinetic cluster analysis can be tested by the rigorous master equation approach for the complete conformational ensemble of the 391 states. By diagonalizing the 391×391 rate matrix, we find that the first four nonzero eigenvalues are 2.087×10^{-10}, 2.27×10^{-6}, 4.88×10^{-6}, and 5.17×10^{-6}. Because there exists a large gap between the first and the second nonzero eigenvalue, the relaxation kinetics is predominantly determined by the slowest mode with a

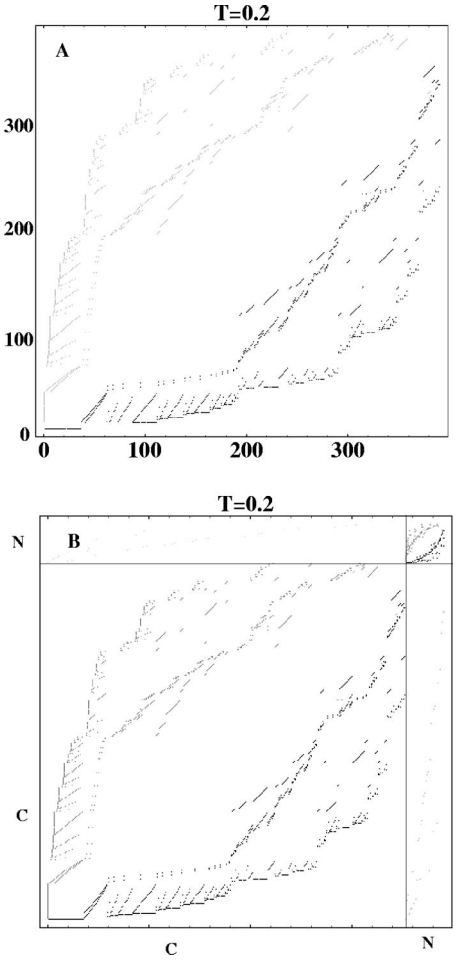

FIG. 6. The transformed block-form rate matrix obtained by sorting conformational states into two distinctive groups (clusters) such that the intracluster transition rates are large and the intercluster transition rates are small (and sparsely distributed).

rate constant of 2.087×10^{-10}, which is in very good agreement with the result $k_R\simeq2.088\times10^{-10}$ computed above from the kinetic cluster analysis.

3. Folding pathways

We assume that the chain is initially in the fully extended conformation. After an initial quick equilibration, the $\Omega_C=353$ conformations within cluster C form a kinetic intermediate before folding to cluster N. There are ten stable conformations each with free energy $G=-5$, which is much lower than the next lowest free energy $G=-4$ with a gap of $5\ k_BT$.

The ten kinetic intermediates contain no native contacts and so they appear as off-pathway kinetic traps in the populational kinetics in Fig. 7. However, since the detrapping from these intermediates are faster than the formation of the rate-limiting stack (5,6,11,12), the ten intermediates are neither off-pathway kinetic traps nor obligatory on-pathway intermediates, because they are not pathway conformations. Therefore, *the disruption of a non-native kinetic intermediate is not necessarily an off-pathway rate-limiting step*.

The most probable pathway is found to be $C_{35}\rightarrow N_{35}$ due to the low free energy (and large fractional population) of state C_{35}. About 95% of the populations fold through this most probable pathway. The most probable folding pathways are dependent on the initial state. For example, the folding for an initial state within the native cluster N would be fast intracluster equilibration.

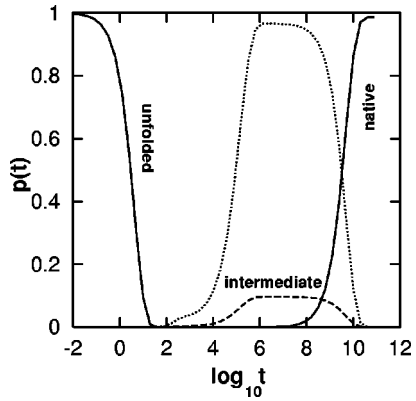

FIG. 7. The population kinetics solved from the exact master equation for the complete conformational ensemble for the folding from the fully extended state. The dashed line is for the population of each of the ten kinetic intermediates, and the dotted line is for the total population of the ten intermediates.

Figure 8 gives the folding, unfolding, and the relaxation rate for a broad temperature range. The kinetic cluster method gives quite accurate results as compared with the results from the exact 391×391 rate matrix.

C. Folding with an off-pathway kinetic trap

We assume that the non-native stack (2,3,6,7) is stabilized by a large enthalpic decrease $\Delta H^*>\Delta H_0$: $(-\Delta S^*, -\Delta H^*)=(-2,-3.4)$. Therefore, the disruption of non-native stack (2,3,6,7) is an off-pathway rate-limiting step due to the large enthalpic barrier ΔH^*. The thermal melting curve shows[43–46] a single peak at the melting temperature $T_m=0.4$. We consider the folding process at $T=0.2<T_m$. At $T=0.2$, the native state [see Fig. 5(A)] occupies a fractional population of 92.3%, and thus the relaxation process is predominantly a folding process. As a result, the relaxation rate is approximately equal to the folding rate k_F.

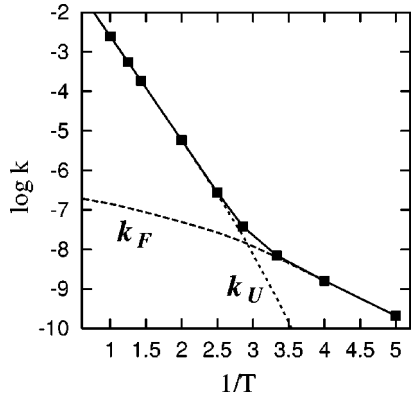

FIG. 8. The temperature dependence of the folding rate k_F (dashed line) and the unfolding rate k_u (dashed line) solved from the kinetic cluster analysis. The solid line and the symbols are for the relaxation rate k_F+k_u solved from the cluster model and from the master equation, respectively.

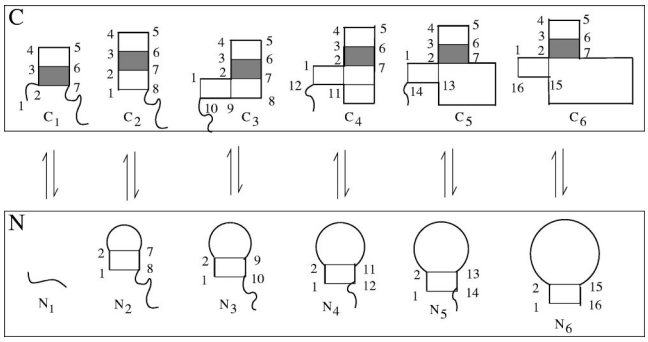

FIG. 9. The detrapping pathways for the six trapped states.

1. Kinetic clusters

The breaking of the non-native stack (2,3,6,7) with a rate constant of $e^{-\Delta H^*/T}=e^{-17}$ is much slower than both the disruption rate $e^{-\Delta H_0/T}=e^{-10}$ and the formation rate $e^{-\Delta S_0}=e^{-5}$ for all the other stacks. Therefore, all conformations with the stack (2,3,6,7) would form a pre-equilibrated cluster C, and the disruption of the stack (2,3,6,7) in these conformations are the rate-limiting steps in the folding process.

We note that the formation rate for the rate-limiting stack $e^{-\Delta S^*}=e^{-2}$ for stack (2,3,6,7) is faster than the disruption and the formation rate for all other stacks. Therefore, it is not appropriate to classify all the conformations without the (2,3,6,7) stack as a pre-equilibrated cluster, because these conformations may form the (2,3,6,7) stack and thus escape the cluster quickly; as a result, conformations without the (2,3,6,7) stack do not form a quasiclosed system and the conditions in Eq. (6) for the formation of a cluster is not satisfied. However, the folding rate, which is essentially equal to the rate for the detrapping from cluster C, is mainly determined by the populational distribution of the trapped conformations in cluster C, not the distribution of the conformations outside cluster C. Therefore, we can compute the folding rate from the inter-''cluster'' transitions: cluster C $\rightarrow$ (all the conformations outside cluster C). In addition, due to the large rate for the formation of the (2,3,6,7) stack, conformations detrapped from cluster C could quickly return to cluster C. Such ''rebounds'' effect would play an important role in the folding kinetics.

2. Detrapping kinetics

We use N to denote the ensemble of conformations outside cluster C, i.e., all the conformations without the stack (2,3,6,7). There are $\Omega_C=6$ conformations in cluster C and $\omega_{C\rightarrow N}=6$ corresponding detrapping pathways; see Fig. 9. Without considering the rebounds effect, the total detrapping rate $k_{C\rightarrow N}$ is determined by Eq. (5) as the sum over the six pathways.

We use the detrapping pathway $C_6\rightarrow N_6$ to illustrate the calculation. The transition $C_6\rightarrow N_6$ has a rate constant of $k_{C_6\rightarrow N_6}=e^{-\Delta H/T}=e^{-17}$ for the breaking of the (2,3,6,7) stack. To account for the rebounds effect, we note that N_6 is connected to 25 kinetically neighboring states through 25 different pathways. Among these 25 kinetically neighboring states, one (C_6) is in the trapping cluster C and the other 24 are outside C. The pathway (state N_6)$\rightarrow$(state C_6) for the formation of the (2,3,6,7) stack is the recrossing transition and has a rate constant of $k_b^{(N_6)}=e^{-\Delta S^*}=e^{-2}$. Among the other 24 pathways, 23 pathways correspond to the different ways to add a stack to N_6 with a rate constant of $e^{-\Delta S_0}$ and one pathway corresponds to the disruption of the (1,2,15,16) stack with a rate constant of $e^{-\Delta H_0/T}$. The total rate for the detrapping through state N_6 is given by the sum for the 24 pathways: $k_f^{(N_6)}=23\cdot e^{-\Delta S_0}+e^{-\Delta H_0/T}=23\cdot e^{-5}+e^{-10}$.

According to Eq. (7), for a folding process starting from state C_6, a fraction of $k_f^{(N_6)}/(k_b^{(N_6)}+k_f^{(N_6)})=53.5\%$ of the population would fold to conformations in N. The remaining 46.5% of the population would recross the barrier and become retrapped in cluster C before refolding to the detrapped ensemble N. Therefore, the effective rate of detrapping from cluster C through route $C_6\rightarrow N_6\rightarrow$ (ensemble N) is given by Eq. (7): $k_{C_6\rightarrow N}^{(\mathrm{eff})}=(P_{C_6})k_{C_6\rightarrow N_6}k_f^{(N_6)}/(k_b^{(N_6)}+k_f^{(N_6)})=4.41\times10^{-9}$. According to Eq. (8), the total rate for $C\rightarrow N$ is determined by the sum for all six trapped conformations C_1-C_6 in Fig. 9: $k_F=\Sigma_{i=1}^{6}k_{C_i\rightarrow N}^{(\mathrm{eff})}=1.07\times10^{-8}$. Under the strong folding condition $T=0.2$, the unfolding rate k_U is small, and the relaxation rate $\simeq k_F=1.07\times10^{-8}$. This result is in close agreement with the exact master equation result, which gives a lowest nonzero eigenvalue of 1.01×10^{-8}.

3. The importance of the rebounds effect and the dominant detrapping pathways

$k_{C_i\rightarrow N}$ (without the rebounds effect) $=0.005\,57$, 0.827, 0.827, 0.827, 0.827, 0.827 ($\times10^{-8}$) for $i=1,2,3,4,5,6$, respectively, and $k_{C_i\rightarrow N}^{(\mathrm{eff})}$ (with the rebounds effect) $=0.003\,54$, $0.000\,277$, 0.0751, 0.213, 0.34, 0.44 ($\times10^{-8}$) for $i=1,2,3,4,5,6$, respectively. The total rate is 4.14×10^{-8} without the rebounds effect and 1.07×10^{-8} with the rebounds effect. The neglect of the rebounds effect indeed causes an significant inaccuracy in predicting the transition rates. In

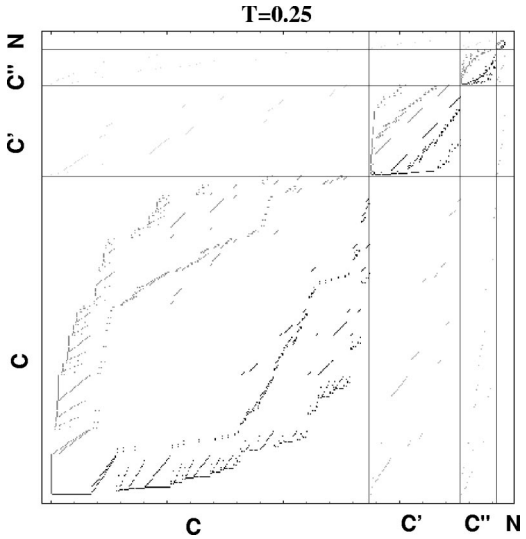

FIG. 10. The transformed block-form rate matrix obtained by sorting conformational states into four distinctive groups (clusters).

addition, from the largest $k_{C_i\to N}^{(\mathrm{eff})}$ values, we find that the dominant contributions for the detrapping of states C_4, C_5 and C_6.

D. Two on-pathway high barriers

In this section, we show how to treat a network of clusters in the presence of multiple rate-limiting steps. We assume that the formation of native stacks (2,3,14,15) and (5,6,11,12) are limited by the large entropic cost ΔS^* ($>\Delta S_0$): $(-\Delta S^*, -\Delta H^*)=(-15,-6)$, so the formation of (2,3,14,15) and (5,6,11,12) are rate-limiting in the folding process. For the given parameters, the melting temperature of the molecule is found[43–46] to be $T_m=0.4$. We consider kinetics at temperature $T=0.25$ ($<T_m$).

1. Kinetic clusters

The conformational space can be classified into the following four clusters: C for the 275 conformations with neither of the two stacks formed, C' for the 78 conformations with only stack (2,3,14,15) formed, C'' for the 30 conformations with only stack (5,6,11,12) formed, and N for the eight conformations with both stacks formed. The native state [see Fig. 5(A)] is in cluster N, and the fully unfolded state is in cluster C. The existence of the four clusters is evident from the block-form rate matrix shown in Fig. 10, such that the intracluster transition rates are notably higher than the (sparse) intercluster transition rates.

2. Intercluster kinetics

The folding process from cluster C to cluster N can be represented by two parallel pathways: $C\leftrightarrow C'\leftrightarrow N$ and $C\leftrightarrow C''\leftrightarrow N$. The folding kinetics can be solved from the master equation for the four-state system (C, C', C'', and N). The key is to compute the 4×4 rate matrix for the transition rates between different clusters.

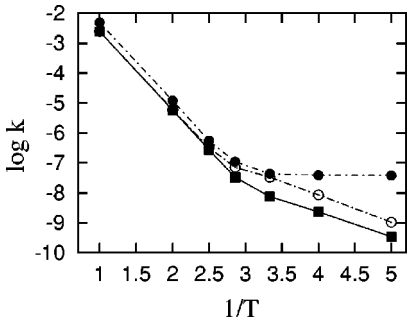

FIG. 11. The lines are for the first three nonzero eigenvalues of the 4×4 rate matrix for the four-cluster system for $T=0.2$–1. The symbols are for the first three nonzero eigenvalues of the original 391×391 rate matrix for the complete conformational ensemble.

There are $\omega_{C\leftrightarrow C''}=26$ pathways connecting conformations in clusters C and C'', each corresponding to the addition and deletion of the (5,6,11,12) with rate $e^{-\Delta S^*}=e^{-15}$ and $e^{-\Delta H^*/T}=e^{-24}$, respectively. Equations (2)–(5) give the following intercluster transition rates: $k_{C\to C''}=6.38\times10^{-10}$ and $k_{C''\to C}=2.02\times10^{-11}$.

Equations (2) and (5) give the following most probable pathways for $C\leftrightarrow C''$: $C_4\to C_2''$; $C_9\to C_6''$; $C_{10}\to C_3''$; $C_5\to C_4''$, each with a rate constant of 1.37×10^{-10}. Therefore, among the 26 pathways, the probability for the molecule to take one of the above four pathways is $1.37\times10^{-10}/6.38\times10^{-10}=21.5\%$, and thus about $4\times21.5\%=86\%$ of the $C\to C''$ transitions are through these four pathways.

Using the similar analysis, we can compute the intercluster rate constants and pathways for the other clusters, and obtain the following 4×4 rate matrix for the four-cluster system:

$$\begin{bmatrix} -8.54\times10^{-9} & 7.90\times10^{-9} & 6.38\times10^{-10} & 0.0 \\ 1.49\times10^{-11} & -2.34\times10^{-9} & 0.0 & 2.33\times10^{-9} \\ 2.02\times10^{-11} & 0.0 & -3.904\times10^{-8} & 3.902\times10^{-8} \\ 0.0 & 1.89\times10^{-12} & 1.89\times10^{-12} & -3.78\times10^{-12} \end{bmatrix},$$

where the row a the order of C, C', C'', and N. The above rate matrix has the following eigenvalues:

$$-\lambda_0=0;\quad -\lambda_1=-2.32\times10^{-9};\quad -\lambda_2=-8.56\times10^{-9};$$

$$-\lambda_3=-3.90\times10^{-8}. \tag{12}$$

From the eigenvector analysis,[17] we find that the kinetic modes for λ_1, λ_2, and λ_3 correspond to the (rate-limiting) transitions $C'\to N$, $C\to C'$, and $C''\to N$, respectively. This is consistent with the following relationships between the eigenvalues and the intercluster rate constants: $\lambda_1\simeq k_{C'\to N}+k_{N\to C'}$, $\lambda_2\simeq k_{C\to C'}+k_{C'\to C}$, and $\lambda_3\simeq k_{C''\to N}+k_{N\to C''}$.

Figure 11 shows that for a wide range of temperatures ($T=0.25$–2.0), the relaxation rates obtained from the kinetic cluster analysis agrees with the eigenvalues for the original 391×391 rate matrix.

3. Folding pathways

Initially, the 275 conformations within cluster C quickly equilibrate and are distributed according to their free energies. As a result, the four most stable states in cluster C would appear as kinetic intermediates in the populational kinetics.

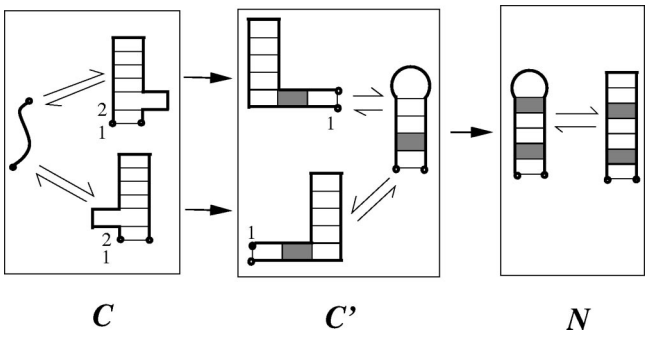

FIG. 12. The most probable folding pathways from C to N.

From the two lowest eigenmodes λ_1 and λ_2, we find that the major folding pathway is $C \rightarrow C'$ (rate$=\lambda_2$) followed by $C' \rightarrow N$ (rate$=\lambda_1$). Furthermore, according to the dominant pathways for $C \rightarrow C'$ and $C' \rightarrow N$, we obtain the two parallel most probable folding pathways as shown in Fig. 12.

A notable feature of the dominant pathways is the absence of the $C \rightarrow C''$ transition. One might expect that two parallel pathways, $C \rightarrow C' \rightarrow N$ and $C \rightarrow C'' \rightarrow N$, would have equal probability because they have exactly the same *total* kinetic barrier $2T\Delta S^*$, where $T\Delta S^*$ is the kinetic barrier for the formation of either stack (2,3,14,15) or stack (5,6,11,12). However, the kinetic partitioning is not only determined by the total barrier along the pathway, but also determined by the distribution of the barriers along the pathways.

Physically, because $k_{C \rightarrow C'} \gg k_{C \rightarrow C''}$, cluster C'' is produced much more slowly than cluster C', and thus the population in cluster C'' does not accumulate significantly. As a result, most of the population would fold along the $C \rightarrow C' \rightarrow N$ pathway. According to intercluster rate constants (see the 4×4 rate matrix above for the four-cluster system), the relative populational partitioning for the folding along $C \rightarrow C'$ and for the folding along $C \rightarrow C''$ is approximately $k_{C \rightarrow C'}/k_{C \rightarrow C''} = 7.90 \times 10^{-9}/6.38 \times 10^{-10} = 12.38$.

From the populational kinetics for each cluster [see Fig. 13(A)] we find a well-populated transient accumulation for cluster C', while there is virtually no populational accumulation for cluster C'' during the folding process. Furthermore, in Fig. 13(B), we show the net flux for transitions $C \rightarrow C'$, $C' \rightarrow N$, $C \rightarrow C''$ and $C'' \rightarrow N$: $P_{C \rightarrow C'}(t) = \int_0^t (P_C(t') \times k_{C \rightarrow C'} - P_{C'}(t') \times k_{C' \rightarrow C}) dt'$, etc. We find that the fluxes

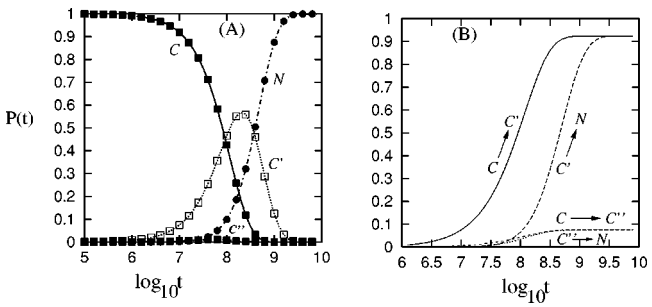

FIG. 13. (A) The population kinetics of each cluster. (B) The net flux for the cluster transitions.

for transitions $C \rightarrow C'$ and $C' \rightarrow N$ are much larger than the fluxes for $C \rightarrow C''$ and $C'' \rightarrow N$. These results confirm that the dominant pathway is $C \rightarrow C' \rightarrow N$.

IV. TEMPERATURE DEPENDENCE OF THE FOLDING RATE

From Eqs. (2)–(5), the intercluster transition rate $k_{C \rightarrow N}$ can be written in the following form:

$$\frac{\sum_{i=1}^{\Omega_{C \leftrightarrow N}} k_{C_i \rightarrow N_i} (P_i/P_{\text{path}})}{1 + (P_{\text{nonpath}}/P_{\text{path}})}, \tag{13}$$

where $k_{C_i \rightarrow N_i}$ is the rate constant for the transition $C_i \rightarrow N_i$, G_{C_i} is the free energy of state C_i, and the ratio

$$P_i/P_{\text{path}} = e^{-G_{C_i}/T} / \sum_{j=1}^{\Omega_{C \leftrightarrow N}} e^{-G_{C_j}/T}, \tag{14}$$

is the fractional population of the ith pathway conformation, and the ratio

$$P_{\text{nonpath}}/P_{\text{path}} = \sum_{j=1+\Omega_{C \leftrightarrow N}}^{\Omega_C} e^{-G_{C_j}/T} / \sum_{j=1}^{\Omega_{C \leftrightarrow N}} e^{-G_{C_j}/T}, \tag{15}$$

is the relative population between the nonpathway and the pathway conformations in cluster C.

From Eq. (13), the temperature dependence of the folding rate k_F comes from the following two factors:

(1) $P_{\text{nonpath}}/P_{\text{path}}$ for the relative population between the pathway and nonpathway conformations. A larger population of the available pathway conformations would give a larger total intercluster folding rate. Moreover, larger energy gap between the pathway and nonpathway conformations leads to a stronger temperature dependence of $P_{\text{nonpath}}/P_{\text{path}}$ and thus a stronger temperature dependence of the folding rate.

(2) The heterogeneity in the population of pathway conformations, as quantified by the fractional distribution P_i/P_{path} for each pathway conformation C_i ($i = 1,2,\ldots,\Omega_{C \leftrightarrow N}$). If different intercluster transitions $C_i \rightarrow N_i$ have the same rate constant $k_{C_i \rightarrow N_i} = k_0$, the temperature dependence due to P_i/P_{path} vanishes because $\sum_{i=1}^{\Omega_{C \leftrightarrow N}} (k_{C_i \rightarrow N_i})(P_i/P_{\text{path}}) = k_0$ is a constant.

In general, different pathway conformations can have different rate constants, e.g., pathway i may have slower rate than pathway j: $k_{C_i \rightarrow N_i} < k_{C_j \rightarrow N_j}$. In such case, if energy E_i of conformation i is higher than energy E_j of conformation j, the relative population between conformation i and j is $P_i/P_j \sim e^{-(E_i - E_j)/T}$ would increase as temperature increases, causing a higher probability to fold through the slow pathway $C_i \rightarrow N_i$, and thus a slower folding rate.

The above two factors are determined by the free-energy landscapes of the molecule. Different free-energy landscapes, e.g., due to different sequences, can have very different temperature dependence of the folding kinetics. Mutations can cause different effects for $P_{\text{nonpath}}/P_{\text{path}}$ and for P_i/P_{path}, and the interplay between these two factors leads

to complex sequence dependence for the temperature dependence of the folding kinetics. For example, mutations can change the folding rate by altering the relative stability and hence the relative distribution ($P_{\text{nonpath}}/P_{\text{path}}$) for the pathway and nonpathway conformations in cluster C. Since the predominant majority of the nonpathway conformations contains non-native stacks, stabilization of the non-native stacks can lead to more populated (misfolded) nonpathway conformations, causing a larger $P_{\text{nonpath}}/P_{\text{path}}$ (and a smaller folding rate k_F). Moreover, the stabilization of the misfolded conformations in cluster C may cause the formation of metastable misfolded kinetic intermediates. The interplay of these two effects would cause a slower folding in general.

Similarly, destabilization of the non-native stacks leads to a smaller population of the nonpathway conformations (i.e., smaller $P_{\text{nonpath}}/P_{\text{path}}$) and thus a faster folding. For example, in the aforementioned folding model with one on-pathway rate-limiting step, if the entropy and enthalpy parameter of a non-native stack is changed from $(-\Delta S_0, -\Delta H_0)=(-5,-2)$ to $(-\Delta S_n, -\Delta H_n)=(-7,-1.40)$, a non-native stack would be destabilized by a free-energy change of $(\Delta H_n - T\Delta S_n)-(\Delta H_0 - T\Delta S_0)=1$ at temperature $T=0.2$ and we expect an accelerated folding. Indeed, our kinetic cluster analysis shows that the folding rate k_F is significantly increased from 2.087×10^{-10} to 7.59×10^{-8}.

The correlation between the folding speed and the relative stability between the native and non-native stacks would help explain and design mutational folding kinetics experiments. In the what follows, based on the 16-nt RNA hairpin model in Fig. 5(A), we investigate the temperature dependence of the folding rate for four mutated model systems.

A. The formation of native stack (6,7,10,11) as a rate-limiting step

Among the 341 conformations in cluster C, there are $\Omega_{C\leftrightarrow N}=50$ pathway conformations. All 50 intercluster pathways are for the formation of the rate-limiting stack (6,7,10,11) with the rate constant $k_0=e^{-\Delta S^*}$. Therefore, according to Eq. (13), the temperature dependence of k_F comes from the factor $P_{\text{nonpath}}/P_{\text{path}}$. Figures 14(I) and 14(II) show the results for two different models.

In Fig. 14(I), we assume $(-\Delta S_0, -\Delta H_0)=(-3.0, -10.0)$ for all the native stacks except for the rate-limiting stack (6,7,10,11) which has $(-\Delta S^*, -\Delta H^*)=(-12.0, -40.0)$, and we assume $(-\Delta S_n, -\Delta H_n)=(-3.0,-3.0)$ for all the non-native stacks. As shown in the figure, as T increases, the nonpathway conformations become more and more stable relative to the pathway conformations, causing a monotonically decreasing folding rate.

Figure 14(II) shows the kinetics for a similar model with a different set of parameters: $(-\Delta S_0, -\Delta H_0)=(-3.0, -6.0)$, $(-\Delta S^*, -\Delta H^*)=(-12.0,-24.0)$, and $(-\Delta S_n, -\Delta H_n)=(-4.0,-5.8)$. The relative stability of the nonpathway conformations first increases then decreases as T is increased, causing a V-shaped k_F versus T curve.

B. The formation of stack (5,6,11,12) as a rate-limiting step

In this case, there exist two clusters N and C for conformational with and without the (5,6,11,12) stack, respectively. There are $\Omega_C=353$ conformations in cluster C and $\Omega_N=38$ conformations in cluster N. There are $\Omega_C=353$ conformations in cluster C, and $\Omega_{C\leftrightarrow N}=35$ of these conformations are pathway conformations. Unlike the model with (6,7,10,11) as the rate-limiting stack, the current model involves inhomogeneous rate constants for different intercluster pathways. There are 32 pathways for the formation of the rate-limiting stack (5,6,11,12). Each of these 32 pathways has a rate constant of $k_1=e^{-\Delta S^*}$. The remaining three pathways correspond to the formation of two consecutive stacks [see Fig. 3(C)]: the (5,6,11,12) stack and a non-rate-limiting native stack. Each of these three pathways has a rate constant of $k_2=e^{-(\Delta S^*+\Delta S)}$. In this case, both $P_{\text{nonpath}}/P_{\text{path}}$ and P_i/P_{path} for each pathway conformation C_i ($i=1,2,\ldots,\Omega_{C\leftrightarrow N}$) would contribute to the temperature dependence of k_F. The combination of $P_{\text{nonpath}}/P_{\text{path}}$ and P_i/P_{path} can give very complex k_F versus T behavior. Figures. 14(III) and (IV) show the results for two different models.

In Fig. 14(III), we assume $(-\Delta S_0, -\Delta H_0)=(-3.0, -10.0)$, $(-\Delta S^*, -\Delta H^*)=(-12.0,-40.0)$, $(-\Delta S_n, -\Delta H_n)=(-3.0,-3.0)$. The factor $P_{\text{nonpath}}/P_{\text{path}}$, shows a V-shape behavior as a function of T, which alone would tend to cause a Λ shape (increasing then decreasing) of the folding rate. However, the P_i/P_{path} factor tends to cause monotonically faster folding for higher temperatures. This is because the conformations C_i on the 32 fast-folding pathways (with rate k_1) become more stable than the slow-folding conformations, so P_i/P_{path} is larger for the fast-folding conformations. The combination of the above two factors leads to a monotonically increasing folding rate.

In Fig. 14(IV), we assume $(-\Delta S_0, -\Delta H_0)=(-8.0, -10.0)$, $(-\Delta S^*, -\Delta H^*)=(-15.0,-40.0)$, $(-\Delta S_n, -\Delta H_n)=(-3.0,-3.0)$. The $P_{\text{nonpath}}/P_{\text{path}}$ factor, which shows an inverted N shape, dominates over the P_i/P_{path} factor, causing a N shape for the k_F versus T curve.

V. APPLICATION TO REALISTIC RNA FOLDING KINETICS

For a realistic RNA hairpin-forming sequence, the clustering is so complex that any method based on simple inspection of the rate constants becomes impossible. The free energies and the rate matrix can be constructed by using the experimentally measured enthalpy and entropy parameters[41] ΔH and ΔS for all possible base stacks. The sequence and the native structure of the hairpin-forming RNA are shown in in Fig. 15(A). There are totally 879 native and non-native structures for this sequence.

After examining the enthalpies and entropies for the formation for all the possible different base stacks, we find that there are two on-pathway rate-limiting steps corresponding to the formation of the native stacks (3,4,18,19) $=(U,C,G,A)$ and (5,6,16,17)$=(G,A,U,C)$, respectively, and two off-pathway rate-limiting steps, corresponding to the disrupting of the non-native stacks (5,6,11,12) $=(G,A,U,C)$, and (11,12,18,19)$=(U,C,G,A)$. According

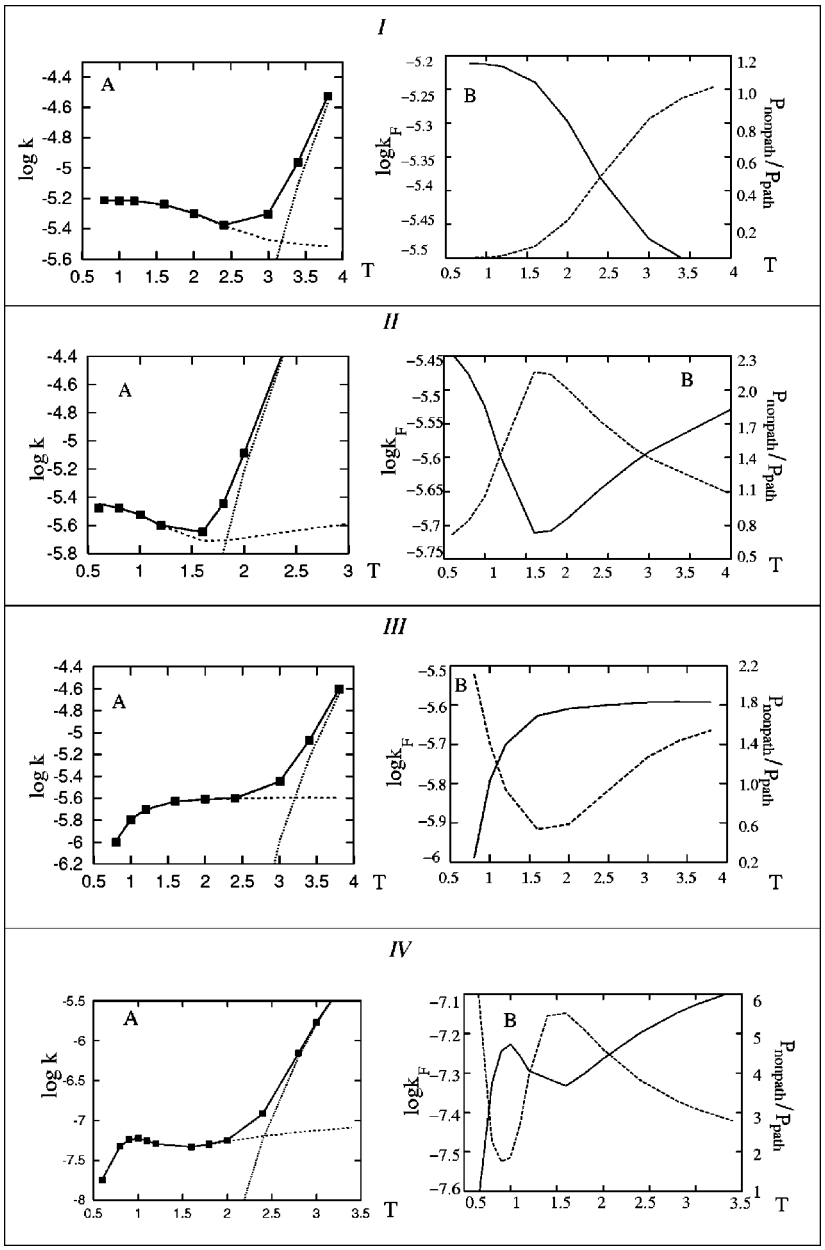

FIG. 14. (A) The temperature dependence of folding rate k_F (dashed line), unfolding rate k_u (dotted line), and the relaxation rate k_F+k_u (solid line) solved from the kinetic cluster method. The filled square symbol represents the lowest nonzero eigenvalue of the original 391×391 rate matrix. (B) The temperature of folding rate k_F (solid line) and the ratio $P_{\text{nonpath}}/P_{\text{path}}$ for the populations of the misfolded state and the on-pathway state (dashed line).

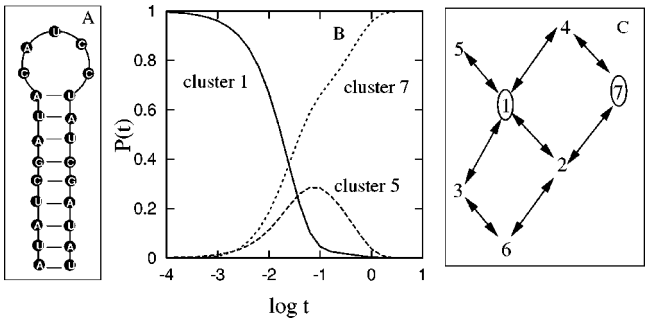

FIG. 15. (A) The native structure for a realistic RNA hairpin-forming sequence. (B) The populational kinetics for a folding reaction at $T=30\,^{\circ}\text{C}$. Kinetics for clusters (cluster 2, 3, 4, and 5) whose fractional population never exceeds 10% are not shown in the figure. (C) The network of kinetic pathways between different clusters. The completely unfolded state is in cluster 1 and the native state is in cluster 7.

to these rate-limiting steps, the conformation space can be classified into the following seven clusters: cluster 1 for the 586 conformations without any of the four stacks formed, cluster 2 for the 105 conformations with stack (3,4,18,19), cluster 3 for the 51 conformations with stack (5,6,11,12), cluster 4 for the 76 conformations with stack (5,6,16,17), cluster 5 for the 36 conformations with stack (11,12,18,19), cluster 6 for the five conformations with both stacks (3,4,18,19) and (5,6,11,12), and cluster 7 for the 20 conformations with both stacks (3,4,18,19) and (5,6,16,17). The native state is in cluster 7, the fully unfolded state is in cluster 1, and there are three misfolded traps: cluster 3, 5 and 6. We consider a folding process starting from the completely unfolded state incluster 1 at $T=30\,^{\circ}\text{C}$. The chain initially undergoes fast equilibration to form the quasiequilibrium cluster 1. After the initial formation of the cluster 1, the chain would fold to the native state which has a fractional popula-

tion of 85% in the final equilibrium state by different pathways. As shown in Fig. 15(C), some of the folding population will directly cross the on-pathway rate-limiting step, (e.g., pathway $1\rightarrow2\rightarrow7$ and $1\rightarrow4\rightarrow7$), and some of the population will undergo trapping and detrapping processes before folding to the native state by crossing the on-pathway rate limiting step (e.g., $1\rightarrow5\rightarrow1\rightarrow4\rightarrow7$, $1\rightarrow3\rightarrow6\rightarrow2\rightarrow7$, etc.)

Using Eqs. (2)–(5), we obtain the 7×7 rate matrix $\mathbf{k}$ for the seven-cluster system

$$\begin{bmatrix} -42.0 & 18.6 & 2.00 & 7.16 & 14.2 & 0.0 & 0.0 \\ 2.01 & -1.023\times10^{3} & 0.0 & 0.0 & 0.0 & 0.213 & 1.021\times10^{3} \\ 29.4 & 0.0 & -58.5 & 0.0 & 0.0 & 29.1 & 0.0 \\ 2.98 & 0.0 & 0.0 & -7.93\times10^{2} & 0.0 & 0.0 & 7.90\times10^{2} \\ 3.66 & 0.0 & 0.0 & 0.0 & -3.66 & 0.0 & 0.0 \\ 0.0 & 29.4 & 29.4 & 0.0 & 0.0 & -58.8 & 0.0 \\ 0.0 & 0.493 & 0.0 & 9.88\times10^{-2} & 0.0 & 0.0 & -0.592 \end{bmatrix},$$

where matrix element k_{ij} $(i\neq j)$ is the rate constant for the transition from cluster i to cluster j. The above rate matrix has the following eigenvalues in the increasing order:

$$\lambda_0=0;\ \lambda_1=2.30;\ \lambda_2=27.6;\ \lambda_3=44.5;$$

$$\lambda_4=88.5;\ \lambda_5=7.93\times10^{2};\ \lambda_6=1.02\times10^{3}. \quad (16)$$

Using the eigenvector analysis,[17] we can identify the kinetic modes for each of the eigenvalues. For example, the lowest nonzero eigenvalue λ_1 corresponds to the intercluster transition $5\rightarrow1$ for the detrapping of the stack (11,12,18,19), λ_2 corresponds to the detrapping of stack (5,6,11,12), and λ_3 corresponds to the formation of the on-pathway rate-limiting stacks (3,4,18,19) and (5,6,16,17). From the rate matrix, we can also estimate fractional population for each pathway starting from cluster 1.

(1) Cluster $1\rightarrow$cluster 5 (misfolded): $k_{15}/(-k_{11})=34\%$, which means about 34% population will be first misfolded into cluster 5 before detrapping (eigenmode λ_1). Moreover, since $k_{15}>k_{51}$, we would expect kinetic accumulation in cluster 5.
(2) Cluster $1\rightarrow$cluster 3 (misfolded): $k_{13}/(-k_{11})=5\%$, which implies that about 5% of the population would first fold into cluster 3.
(3) Cluster $1\rightarrow$cluster 2 (on-pathway) and cluster $1\rightarrow$cluster 4 (on-pathway): $(k_{12}+k_{14})/(-k_{11})=61\%$. Since cluster 2 is kinetically connected to the misfolded trapping clusters 3 and 6, the fractional population for the folding without being trapped would be less than 61%. And the fraction population for the folding through the trapping–detrapping processes would be more than 5%. Moreover, because $k_{27}\gg k_{12}$ and $k_{47}\gg k_{14}$, the fraction of population in clusters 2 and 4 would quickly fold into the native cluster 7. As a result, there is no significant kinetic accumulation of population in clusters 2 and 4.

According to Eqs. (16) and (1), we obtain the population for the native cluster 7: $P_{\text{cluster7}}(t)=P^{\text{eq}}_{\text{cluster7}}-0.40e^{-\lambda_1 t}-0.156e^{-\lambda_2 t}-0.462e^{-\lambda_3 t}$, where $P^{\text{eq}}_{\text{cluster7}}=0.999$ is equilibrium population of cluster 7. Because the fractional population of the native state in cluster 7 is $P^{\text{eq}}_{\text{native state}}/P^{\text{eq}}_{\text{cluster7}}=0.85$, we obtain the following native populational kinetics: $P^{\text{eq}}_{\text{native state}}=0.85-0.33e^{-\lambda_1 t}-0.13e^{-\lambda_2 t}-0.39e^{-\lambda_3 t}$. In Fig. 15(B), we plot the populational kinetics curves for the native state and for all the 7 clusters. Indeed, we find that cluster 5 is a kinetic intermediates while cluster 2 and 4 do not show significant kinetic accumulation, which agrees well with the analysis.

To validate our kinetic cluster analysis, we have also solved the eigenmodes for the original 879×879 rate matrix for the complete 879 conformational states. The first seven eigenvalues are

$$\lambda'_0=0;\ \lambda'_1=2.27;\ \lambda'_2=27.5;\ \lambda'_3=43.8;$$

$$\lambda'_4=88.1;\ \lambda'_5=7.92\times10^{2};\ \lambda'_6=1.02\times10^{3}. \quad (17)$$

We find that the first seven eigenvalues of the 879×879 rate matrix agree nearly exactly with the above eigenvalues for the 7×7 rate matrix for the intercluster kinetics. Furthermore, the native populational kinetics solved from the original 879×879 rate matrix is given by $P^{\text{eq}}_{\text{native state}}=0.85-0.34e^{-\lambda'_1 t}-0.13e^{-\lambda'_2 t}-0.38e^{-\lambda'_3 t}$, which also agrees very well with results from the cluster analysis.

VI. DISCUSSIONS

We have developed a kinetic cluster method to analyze the folding rates and folding pathways. The method is based on the classification of the conformational ensemble into clusters. Different clusters are separated by high kinetic barriers, and thus conformation in each cluster can preequilibrate before crossing the intercluster barriers. In terms of the clusters, the overall kinetic process can be represented as the intercluster transitions.

Our intercluster transition rate calculation accounts for the effect of multiple pathways and the effect of possible local minima on the free-energy landscape between the clusters. We are able to identify the dominant pathways from the intercluster kinetic analysis. In addition, we found that for nontrapping local minima, the increase of the stability may accelerate the folding process.

Conformations can be classified into different clusters in different temperature regimes. For example, if the formation of a native stack n causes a large entropic loss ΔS_n^*, conformations without the rate-limiting stack n would form a pre-equilibrated cluster C. If there exists a non-native stack nn that requires a large enthalpic cost ΔH_{nn}^* for disruption, then for low temperatures $T<\Delta H_{nn}^*/\Delta S_n^*$, the barrier ΔH_{nn}^* (for the breaking of nn) would exceeds the barrier $T\Delta S_n^*$ (for the formation of n), and thus the disruption of the misfolded conformations with stack nn is slower than the formation of the rate-limiting native stack n. As a result, conformations with stack nn must be separated out from cluster C to form a separate cluster. However, for temperatures $T>\Delta H_{nn}^*/\Delta S_n^*$, the misfolded conformations with stack nn can quickly equilibrate with other conformations in cluster C, and thus need not be treated as a separate cluster.

Moreover, within a given temperature regime, the conformations can be classified as the same set of clusters, but the folding rate can be quite different for different temperatures. Based on the kinetic cluster analysis, we are able to analyze the temperature dependence of the folding and unfolding rate from the relative stability between the pathway conformations and the nonpathway conformations and that between different pathway conformations. The ability to predict and to analyze the temperature dependence of the folding rate would greatly enable us to design sequences with specific temperature dependence of the folding rate.

For all the sequences and energy landscapes that we have investigated, we found that the quasiequilibrium condition was satisfied for all the conformations within the clusters. We assume that the intracluster transitions are fast so that fast equilibration between conformations can be realized. For conformations that cannot be directly converted through a single kinetic move, we assume that there exist fast routes through multiple (fast) kinetic moves to connect these conformations. In fact, such fast intracluster routes have been identified for a number of conformations that we have examined.

The present theory has been illustrated by the computation with the simple hairpin-forming molecules. For the purpose of illustrating the principle of the method, we used simplified models, but the kinetic cluster theory developed here is general and can be developed to treat the folding problems of complex biopolymers with realistic chain length.

ACKNOWLEDGMENTS

This work has been supported by grants from NIH (GM063732) and from AHA National Center (0130064N).

[1] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).
[2] K. A. Dill and H. S. Chan, Nat. Struct. Biol. **4**, 10 (1997).
[3] V. S. Pande and D. S. Rokhsar, Proc. Natl. Acad. Sci. U.S.A. **96**, 9062 (1999).
[4] D. K. Klimov and D. Thirumalai, Proteins **43**, 465 (2001).
[5] H. Isambert and E. D. Siggia, Proc. Natl. Acad. Sci. U.S.A. **97**, 6515 (2000).
[6] D. Thirumalai and S. A. Woodson, Acc. Chem. Res. **29**, 433 (1996).
[7] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).
[8] R. Zwanzig, Proc. Natl. Acad. Sci. U.S.A. **94**, 148 (1997).
[9] M. Cieplak, M. Henkel, J. Karbowski, and J. R. Banavar, Phys. Rev. Lett. **80**, 3654 (1998).
[10] Y.-J. Ye, D. R. Ripoll, and H. A. Scheraga, Comput. Theor. Polym. Sci. **9**, 359 (1999).
[11] Y.-J. Ye and H. A. Scheraga, in *Slow Dynamics in Complex Systems*, AIP Conference Proceedings, No. 469, 1999, p. 452.
[12] M.-H. Hao and H. A. Scheraga, J. Phys. Chem. **107**, 8089 (1997).
[13] S. B. Ozkan, I. Bahar, and K. A. Dill, Nat. Struct. Biol. **8**, 765 (2001).
[14] W. B. Zhang and S. J. Chen, Proc. Natl. Acad. Sci. U.S.A. **99**, 1931 (2002).
[15] M. Tacker, W. Fontana, P. F. Stadler, and P. Schuster, Eur. Biophys. J. **23**, 29 (1994).
[16] V. Munoz and W. A. Eaton, Proc. Natl. Acad. Sci. U.S.A. **96**, 11311 (1999).
[17] W. B. Zhang and S. J. Chen, J. Chem. Phys. **118**, 3413 (2003).
[18] J. D. Bryngelson and P. G. Wolynes, J. Phys. Chem. **93**, 6902 (1989).
[19] S. B. Ozkan, K. A. Dill, and I. Bahar, Biopolymers **68**, 35 (2003).
[20] R. Czerminski and R. Elber, J. Chem. Phys. **92**, 5580 (1990).
[21] O. M. Becker and M. Karplus, J. Chem. Phys. **106**, 1495 (1997).
[22] S. Krivov and M. Karplus, J. Chem. Phys. **117**, 10894 (2002).
[23] S. C. Smith, J. Phys. Chem. **104**, 10489 (2000).
[24] A. Ghosh, R. Elber, and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **99**, 10394 (2002).
[25] H. S. Chan and K. A. Dill, J. Chem. Phys. **100**, 9238 (1994).
[26] H. S. Chan and K. A. Dill, J. Chem. Phys. **99**, 2116 (1993).
[27] Y. Konishi, T. Ooi, and H. A. Scheraga, Biochemistry **21**, 4734 (1982).
[28] Y. Konishi and H. A. Scheraga, Biochemistry **19**, 1308 (1980).
[29] Y. Konishi and H. A. Scheraga, Biochemistry **19**, 1316 (1980).
[30] Y. Konishi, T. Ooi, and H. A. Scheraga, Biochemistry **20**, 3945 (1981).
[31] Y. Konishi, T. Ooi, and H. A. Scheraga, Biochemistry **20**, 4741 (1982).
[32] H. S. Chan and K. A. Dill, Proteins: Struct., Funct., Genet. **30**, 2 (1998).
[33] H. Eyring, J. Chem. Phys. **3**, 107 (1935).
[34] N. S. Snider, J. Chem. Phys. **42**, 548 (1964).
[35] B. Widom, J. Chem. Phys. **55**, 44 (1971).
[36] G. J. Moro, J. Chem. Phys. **103**, 7514 (1995).
[37] D. Shalloway, J. Chem. Phys. **105**, 9986 (1996).
[38] D. Shalloway, J. Chem. Phys. **109**, 1670 (1998).
[39] T. L. Hill, *An Introduction to Statistical Thermodynamics*, Addison–Wesley Series in Chemistry (Addison–Wesley, Reading, MA, 1960).
[40] C. Wagner and T. Kiefhaber, Proc. Natl. Acad. Sci. U.S.A. **96**, 6716 (1997).
[41] M. J. Serra and D. H. Turner, Methods Enzymol. **259**, 242 (1995).
[42] M. Zuker, Science **224**, 48 (1989).
[43] S.-J. Chen and K. A. Dill, J. Chem. Phys. **109**, 4602 (1998).
[44] S.-J. Chen and K. A. Dill, J. Chem. Phys. **103**, 5802 (1995).
[45] S.-J. Chen and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **97**, 646 (2000).
[46] W. B. Zhang and S. J. Chen, J. Chem. Phys. **114**, 7669 (2001).