S. Banu Ozkan[1,2]
Ken A. Dill[3]
Ivet Bahar[1,2]

[1] Center for Computational
Biology & Bioinformatics,
and Department of Molecular
Genetics & Biochemistry,
School of Medicine,
University of Pittsburgh,
PA 15213, USA

[2] Department of Chemical
Engineering & Polymer
Research Center,
Bogazici University,
Bebek 80815,
Istanbul, Turkey

[3] Department of
Pharmaceutical Chemistry,
University of California
San Francisco,
CA, 94143-1204, USA

Computing the Transition State Populations in Simple Protein Models

Received 24 June 2002;
accepted 12 August 2002

Abstract: *We describe the master equation method for computing the kinetics of protein folding. We illustrate the method using a simple Go model. Presently most models of two-state fast-folding protein folding kinetics invoke the classical idea of a transition state to explain why there is a single exponential decay in time. However, if proteins fold via funnel-shaped energy landscapes, as predicted by many theoretical studies, then it raises the question of what is the transition state. Is it a specific structure, or a small ensemble of structures, as is expected from classical transition state theory? Or is it more like the denatured states of proteins, a very broad ensemble? The answer that is usually obtained depends on the assumptions made about the transition state. The present method is a rigorous way to find transition states, without assumptions or approximations, even for very nonclassical shapes of energy landscapes. We illustrate the method here, showing how the transition states in two-state protein folding can be very broad ensembles.* © 2002 Wiley Periodicals, Inc. Biopolymers 68: 35–46, 2003

Keywords: *master equation method; kinetics of protein folding; Go model; two-state fast-folding; transition state*

INTRODUCTION

Proteins fold or unfold via kinetic processes that are usually well described as sums of exponentials. Small proteins often fold rapidly with the simplest possible kinetics: a single exponential relaxation in both folding and unfolding directions. Substantial theoretical and experimental efforts have been devoted for defining the protein conformation(s) that correspond to the slowest exponential relaxation rates—the rate-limit-

Correspondence to: Ivet Bahar, Center for Computational Biology and Bioinformatics, University of Pittsburgh School of Medicine, L. Kaufmann Building, Suite 601, 3471 Fifth Avenue, Pittsburgh, PA 15213; email: bahar@pitt.edu

Biopolymers, Vol. 68, 35–46 (2003)

ing steps. The problem has been that most theoretical treatments require some critical ad hoc assumptions; hence the predicted transition states (TS) may be the reflections of the flawed assumptions about what the TS is. In classical chemical reaction theory, there are usually well-defined single structures (or small ensembles) that correspond to the reactant, the TS, and the product so the task of identifying a structure responsible for the rate-limiting step succeeds. However, for proteins, while the native state is unique, the denatured state is a broad ensemble of conformations, raising the question of how to characterize the transition state—as unique structures, or as ensembles. Here we provide a rigorous way to identify the transition state populations in models of any complexity. We describe the kinetics using a master equation.

The master equation formalism[1–6] has been developed for protein folding kinetics and applied in a number of earlier studies. Lepold et al. studied the folding of simple lattice chains having different sequences by using the master equation approach.[7] In their study, the only transitions allowed were local conformational changes. They concluded that the foldability of a sequence is predicted if there is a single folding funnel leading to a native state and nonfoldability is predicted if multiple pathways lead to several stable conformational states. A transition matrix approach that is equivalent to a master equation formalism with a finite time approximation was used by Chan and Dill[8] for analyzing macromolecular collapse dynamics. Their method is based on considering the transition probabilities at certain time intervals. As the time unit becomes infinitesimally small, the approach reduces to a standard master equation formalism. Chan and Dill enumerated all the conformations of a simple lattice model and applied the transition matrix approach to explore all the possible kinetic pathways for the folding of the simple lattice model.

Many proteins fold via two-state kinetics. Two-state kinetics applies if the protein molecules equilibrate rapidly between different unfolded conformations prior to complete folding. Zwanzig was the first to apply the master equation formalism to describe protein folding by a two-state kinetics.[9,10] Ye et al. used the general Laplace transformation solution of master equation formalism to describe the folding kinetics of a small portion of staphyloccocal protein A.[11] They concluded that the protein folds in a fast cooperative process, and that neither the initial state nor the number of local energy minima affect the long time kinetics of folding.

The master equation for 12-monomer on-lattice heteropolymers has been solved numerically by Cieplak et al. and the time evolution of the occupancy of the native state has been determined.[12] In order to understand the mechanism of secondary structure formation, Eaton and co-workers studied the formation of a β-hairpin using a master equation formalism.[13] Zhang and Chen analyzed RNA folding kinetics using the same methodology.[14] We recently used the master equation formalism with simple lattice model chains for gaining an understanding of the physical basis of unusual Φ values observed in folding kinetics experiments.[15] Here we present the general master equation approach, applicable for any folding model, and we show how to use it to unambiguously identify the "transition state" for folding, even if the system does not have the classical type of single bottleneck step.

FORMULATION OF THE EQUATION

Stochastic processes underlie much of physics, chemistry [6] and biology,[16,17] including population dynamics and epidemiology. In the present study, we analyze the stochastic process of a protein folding described as an ensemble of transitions between N accessible conformational states. If $N = 2$, this is the two-state classical mass-action model, but larger N's are appropriate for more microscopic models in which the non-native states are listed exhaustively (only possible for very simple models) or are collected together in some meaningful way as "macrostates." The time evolution of these states is controlled by the *master equation*,

$$d\mathbf{P}(t)/dt = \mathbf{A}\mathbf{P}(t) \tag{1}$$

where $\mathbf{P}(t)$ is the N-dimensional vector of the instantaneous probabilities of the N conformations, and $\mathbf{A}$ is the N × N *transition (*or *rate) matrix* describing the kinetics of the transitions between these conformations. By definition, the ijth off-diagonal element (A_{ij}) of $\mathbf{A}$ is the rate coefficient for the passage from conformation j into conformation i. From the principle of detailed balance, $A_{ij}\, p_j^0 = A_{ji}\, p_i^0$, where p_i^0 is the equilibrium probability of the ith conformation. The ith diagonal element of $\mathbf{A}$ represents the overall rate of escape from conformation i. It is found from the negative sum of the off-diagonal elements in the same column, i.e., $A_{ii} = -\Sigma_j A_{ji}$ $(j \neq i)$.

Equation (1) represents a set of N simultaneous ordinary differential equations. The formal solution can be cast into a tractable form by decomposing $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{B}\Lambda\mathbf{B}^{-1} \tag{2}$$

where **B** is the matrix of the eigenvectors of **A**, **Λ** is the diagonal matrix of its eigenvalues λ_i ($\lambda_1 = 0$ and $\lambda_i < 0$ for $2 \leq i \leq N$) and $\mathbf{B}^{-1}$ is the inverse of **B**. The time-dependent probability of occurrence of the *i*th conformation [i.e., the *i*th element of **P**(*t*)] can be expressed in terms of the elements of **B**, **Λ**, $\mathbf{B}^{-1}$, and **P**(0) as

$$P_i(t) = \sum_{j=1}^{N} \sum_{k=1}^{N} B_{ik} \exp(\lambda_k t)[B^{-1}]_{kj} P_j(0)$$

$$= \sum_{j=1}^{N} C(i, t|j, 0) P_j(0) \quad (3)$$

where $C(i, t \mid j, 0)$, denotes the *conditional* or *transition probability* of conformation *i* at time *t*, given that the chain was in conformation *j* at $t = 0$. For stationary processes, $C(i,t|j,0)$ is independent of the initial time of observation, but depends only on the time interval *t* between two successive conformations; that is, $C(i,t_2|j,t_1) = C(i,t|j,0)$ for $t = t_2 - t_1$. In matrix notation, Eq. (3) reads

$$\mathbf{P}(t) = \mathbf{B} \exp\{\Lambda t\} \mathbf{B}^{-1} \mathbf{P}(0) = \mathbf{C}(t) \mathbf{P}(0) \quad (4)$$

where $\exp\{\mathbf{\Lambda} t\}$ is a diagonal matrix whose *i*th element is $\exp\{\lambda_i t\}$, and **C**(*t*) is the conditional or transition probability matrix. **C**(*t*) fully describes the stochastic process of $N \times N$ transitions. The *time-delayed joint probability* of conformations *i* at time t_2 and *j* at time t_1 is found from

$$P(i, t_2; j, t_1) = C(i, t_2 - t_1|j, 0) P_j(t_1) \quad (5)$$

Combining these probabilities, we obtain the time-delayed joint probability,

$$P(A, t_2; B, t_1) = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} C(i, t_2 - t_1|j, 0) P_j(t_1) \quad (6)$$

for specific conformational subsets (or macroconformations) *A* and *B* of interest. N_A and N_B denote the numbers of conformations in these subsets.

Rearranging Equation (3) gives $P_i(t)$ as a sum of exponentials

$$P_i(t) = \sum_{k=1}^{N} a_{ik} \exp(\lambda_k t) \quad (7)$$

where $-\lambda_k$ is the frequency of the *k*th mode of motion, and a_{ik} is the corresponding amplitude factor. The a_k is an equilibrium property characteristic of the *k*th mode; it is related to the eigenvectors of **A** and the initial distribution of conformations as

$$a_{ik} = \sum_{j=1}^{N} B_{ik}[B^{-1}]_{kj} P_j(0) \quad (8)$$

This equation follows from comparison of Equations (3) and (7). The frequencies are usually organized in ascending order, such that $\lambda_1 = 0$, and $-\lambda_2$ is the frequency of the slowest mode of conformational motion. The latter describes the *global* folding mode, while the high frequency modes refer to *local* structure formation or conformational fluctuations. In short, whenever the eigenvalue spectrum separates into many fast modes, and a single slowest mode, $-\lambda_2$, as it always does for two-state protein folders, the fast modes correspond to a "burst phase," and the slow mode corresponds to the kinetic process that is attributed to a "transition state," the single slowest rate of the process.

MODELS AND NATIVE CONFORMATIONS

In the present study, we illustrate the method using short model chains (9-mers and 16-mers) on square lattices. These models exhibit a two-state kinetics according to the criterion that the time-dependent formation (or accumulation) of the native state is well approximated by a single-exponential decay in time. The reason for illustrating with these simple models is that they are the only models that can be studied without approximation, and therefore they best illustrate the principles and generality of the method. The method gives the time evolution of the complete ensemble of $N \times N$ conformational transitions, thus providing exact and detailed information on the mechanism or pathway(s) of folding. The accessible conformations consist of all self-avoiding walks generated on a square lattice, including both the extended conformations that dominate the denatured state, and compact forms confined to 3×3 (or 4×4) lattices. Exhaustive enumeration yields $N = 740$ and 802,075 distinct conformations for the 9-mers and 16-mers, respectively, excluding the conformers that are related by symmetry or rigid body rotation. The analysis of the complete ensemble enables us to capture the microscopic detail, the sequence–kinetics relationship,

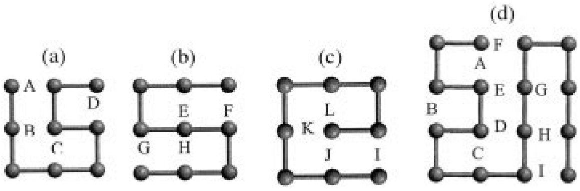


FIGURE 1 Three maximally compact conformations for the 9-mer on 3 × 3 square lattice, shown in parts (a)–(c). Each have four intramolecular contacts, labeled as *A–D*, *E–G*, and *I–L*, respectively. Part (d) shows the selected native structure (among the total of 31 maximally compact structures) for the 16-mer. This structure may be viewed as a simplified model for a protein comprising two domains, an *α-helical* (*A–C*) and a *β-sheet* domain (*G–I*), while *D–F* are *interdomain* contacts.

or the structural aspects of how the chains actually fold.

Native Conformation

There exist three maximally compact conformations for a 9-mer on a square lattice [Figure 1(a)–(c)]. Each have four intramolecular contacts, labeled as A–D, E–G, and I–L, respectively. Calculations were performed for each of these conformations selected as the native state. The conformation (a) is examined in more detail, since this model involves both local (between monomers i and $i + 3$) and nonlocal contacts, grouped in two sequentially separate domains.

The 16-mer, on the other hand, has 31 maximally compact conformations having nine contacts each (confined to the 4 × 4 lattice). Among these, the conformation shown in Figure 1(d) is selected in the present study as the native structure. This structure may be viewed as a simplified model for a protein comprising two domains, an α-helical and a β-sheet. Contacts A–C are representative of helical contacts, and G–I are β-strand contacts. These two sets of three contacts may be viewed as local and nonlocal *intradomain* contacts, respectively, while D–F are *interdomain* contacts that assemble these secondary structures. The analysis of the time evolution of these contacts should provide insights about the hierarchical formation, if any, of different types of contacts during the folding process.

ENERGETICS AND PARAMETERS

We study the passage from a broad distribution of conformations (the denatured state) to a well-defined native conformation using this master equation method. For purpose of illustration, we use a Go model.[25] Folding is driven by attractive potentials assigned to pairs of monomers making native contacts. Forming a native contact involves an energy decrease of ϵ. Dissociating the contact increases the energy by ϵ; all non-native contacts have zero energy. It is known that non-native contacts may contribute to stability.[18] However, Go models are used because they have the same large conformational search as proteins; they have a unique lowest energy "native" state, and they exhibit two-state kinetics. The energy landscape is not yet known for atomistic models but it is possible to explore the landscape using a Go model. Go models have proven useful in earlier folding kinetics studies,[19,20] and their results were comparable to those obtained with full models.[19–21]

Conformational transitions are assigned rate constants based on intramolecular energy barriers and friction. The energy barrier height is taken as zero for passages to a conformation of equal or lower energy, and as the difference in energy between the initial and final conformations when passing to a higher energy conformation. The frictional effect accounts for the geometric accessibility of one conformation from another. Transitions are slow between conformations that are very dissimilar, and fast between similar conformations. The frictional term in our rate matrix depends on the root mean square (rms) deviation between the bead positions of the two conformations. The rate constant A_{ij} for the passage between conformations j and i is given by

$$A_{ij} = \exp\{-\Delta G_{ij}/RT\} = \exp\{-\nu\langle(\Delta r_{ij})^2\rangle^{1/2}\} \times \exp\{-(q_i - q_j)\varepsilon H(q_i, q_j)/RT\} \quad (9)$$

where ΔG_{ij} is the free energy change accompanying the transition, q_i is the number of native contacts occurring in conformation i, $H(q_i, q_j)$ is the Heavyside step function, equal to one for $q_j > q_i$, and zero otherwise. The $\langle(\Delta r_{ij})^2\rangle^{1/2}$ is the rms deviation between the conformations i and j evaluated after optimal superposition of the two conformations, and ν is a proportionality constant dependent on frictional effect. In the absence of viscous effects, $\nu = 0$. An alternative model would be an inverse dependence on macroscopic viscosity, following Kramer's rate expression, in conformity with the modeling of protein folding as a diffusion process.[22] But here we prefer the more microscopic strategy, since it discriminates between individual transitions on the basis of their three-dimensional structures. Figure 2 illustrates the time evolution of the native conformation at different ν values. As the frictional resistance increases, there is an increase in the time elapsed for reaching the native state; yet, the equilibrium distribution of conforma-

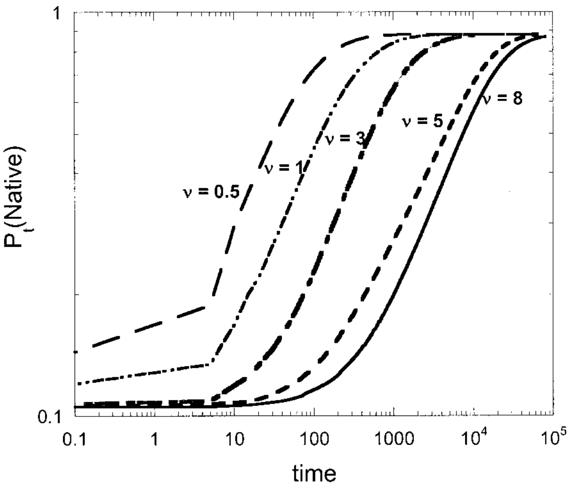


FIGURE 2 Time evolution of native conformation at different friction constants ν. Same equilibrium is reached in all cases with the same Go potentials; the increase in ν simply induces a time lag in reaching the equilibrium state.

tions is unaffected provided that the same Go potential parameters are used. Here, the values $\epsilon = -5$ *RT* and $\gamma = 0.5$ were adopted for the 9-mers, and $\varepsilon = -2.3$ *RT* and $\nu = 1.0$ for the 16-mers. Bonds have unit length. The relatively weaker potentials and higher frictional resistances used in the 16-mers result from physical and technical reasons: (a) physically, the moderate driving potential for folding enables us to examine the time evolution of contacts and possible accumulation of intermediates when the folding kinetics becomes more complex; (b) technically, computational overflows are avoided, which would otherwise arise from the exceedingly broad difference in the time scale of the fast and slow transitions.

INITIAL CONDITIONS, EQUILIBRIUM DISTRIBUTION AND TIME STEPS

Calculations for the 9-mers are performed using a uniform distribution of all conformations, i.e., $P_i(0) = 1/N = 1/740$ for all i as initial conditions. This represents the infinite temperature limit. The ensemble converges to the Boltzmann distribution at 300 K at long times. The equilibrium probability of the native conformation (n) is $P_n(\infty) = 0.9848$ using $\varepsilon = -5$ *RT*. Therefore, the stochastic process of folding to the native state starting from a uniform distribution of conformations is observed to investigate the different pathways.

In the case of the 16-mers, we use the Boltzmann distribution at 500 K as the initial distribution, rather than, say, the infinite temperature distribution. The choice is of little consequence. The net effect is to reduce the high probability of conformations having no contacts or one contact at the initial stage of folding. The equilibrium probability of the native conformation is 0.008 at $T = 500$ K, and 0.837 at $T = 300$ K using $\epsilon = -2.3$ *RT*. The equilibrium probability of the energetically nearest conformation making eight native contacts instead of nine, contact I at chain terminus being disrupted, is 0.086 at $T = 300$ K. Thus the total equilibrium probability of the two conformations account for more than 92% of the observed molecules at equilibrium.

The master equation formalism permits us to explore the folding processes over the full range of time scales. Time steps Δt of different sizes can be used, depending on the time scale or the stochastic process of interest. Steps of $\Delta t = 0.01$ time units were adopted, for example, for examining the initial folding stages in the 9-mers, while the later stages were examined with 3–4 orders of magnitude larger time steps, consistent with the observed distributions of frequencies (eigenvalues of **A**). A broader distribution spanning about five orders of magnitude is operated in the case of 16-mers. This way, it is possible to observe both the local structure formation and global folding processes using the same methodology, which is otherwise impossible in detailed-model simulations by Monte Carlo or molecular dynamics.

KINETICS IN TERMS OF MACROCONFORMATIONS

We analyzed our results in terms of subsets of conformations, in order to reach a macroscopic description of folding mechanisms, as experimentalists prefer. In the case of the 9-mer, the conformations are grouped into 13 subsets, according to their number and types of native contacts: Subset O comprises the conformations having no native contact; subsets A, B, C, and D contain those having only one (A, B, C, or D) native contact each, as indicated by their name; AB, AC, BC, and BD have two native contacts; ABC and BCD have three contacts; and finally ABCD is the native conformation (having four contacts). We note that a number of macroconformations such as AD, ABD, and ACD are not accessible to the lattice geometry.

The important question that is addressed is whether a reduced 13 × 13 model of macroconformations can accurately describe the kinetic process extracted from the 740 × 740 matrix of microconformations for the 9-mer. The reduced model is much faster to compute. Reducing the size of **P**(t) by one order of magnitude indeed increases the computational efficiency by two

Table I The Total Number of Microconformations (W_{mic}) and Macroconformations (W_{mac}) for the Conformations Having the Same Number of Native Contacts (m)

m	W_{mic}	W_{mac}
1	258453	9
2	81992	35
3	21024	68
4	3889	76
5	522	50
6	94	20
7	14	6
8	1	1

orders of magnitude, i.e., the computation time scales with N^2. The examination of the reduced set of macroconformations enables us to extend the methodology to longer chain models, whose computations are otherwise prohibitively expensive. Furthermore, experimental data are usually interpreted in terms of such ensemble averages, so the reduced model is a closer descriptor of experimental results.

Based on these arguments, we construct a reduced 13 × 13 transition matrix $\mathbf{A}'$, the elements of which describe the rates of passages between the macroconformations. The element accounting for the passage from macroconformation B to A, for example, is found by double summing the elements A_{ij} of $\mathbf{A}$ over $1 \leq i \leq N_A$ and $1 \leq j \leq N_B$ [see Eq. (6)]. Calculations repeated for the reduced (13-d) probability array showed that the time evolutions of the native structure and the native contacts are identical to those obtained with the 740-d arrays. Hence reduction to macrostates is warranted.

The calculations for the 16-mer were likewise carried out in a reduced space of 257 macroconformation, which includes all possible distributions of native contacts except for ten having ≤1 native contacts. Table I lists the total number of macroconformations and microconformations for 16-meric structures having the same number of native contacts.

The rms deviation $\langle(\Delta r_{ij})^2\rangle^{1/2}$ between macroconformations i and j were found on the basis of the average radii of gyration of the conformations belonging to the two macroconformations. This approximation is tested with the 9-mers and verified to lead to insignificant changes in the observed results.

DISPERSION OF TRANSITION MODES OF THE MACROCONFORMATIONS

The eigenvalue analysis of the reduced transition rate matrix $\mathbf{A}'$ allows us to visualize the type and relative time scale of the individual modes of motion that contribute to the folding process. The eigenvalues λ_k' ($2 \leq k \leq 13$) of $\mathbf{A}'$ are representative of the frequencies of the modes. And the associated eigenvectors describe the shapes of the individual modes.

The decomposition of the transition rate matrix $\mathbf{A}$ of the 9-mer gives a trimodal distribution that is presented in Figure 3. The trimodal distribution could be classified as (a) a burst stage at $t < 0.03$ time units, approximately, (b) intermediate times $0.03 \leq t < 2$ time units, and (c) long times $t \geq 2$ time units. At the burst stage, only the fastest modes operate. This is a rapid decay of the conformations having no native contacts. At long times, on the other hand, only one slowest mode contributes; this mode dominates the

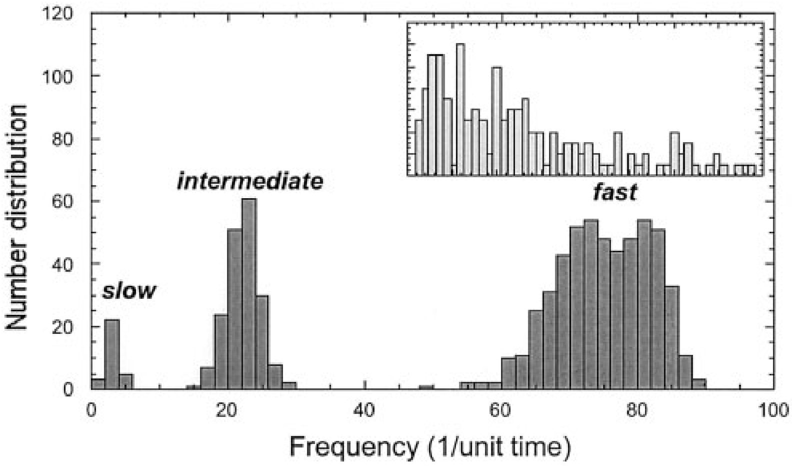


FIGURE 3 Trimodal distribution of frequencies for 9-mer, found by eigenvalue decomposition of the rate matrix **A**. Three time regimes—burst, intermediate, and long times are distinguishable. The inset presents the frequency distribution for the 16-mer.

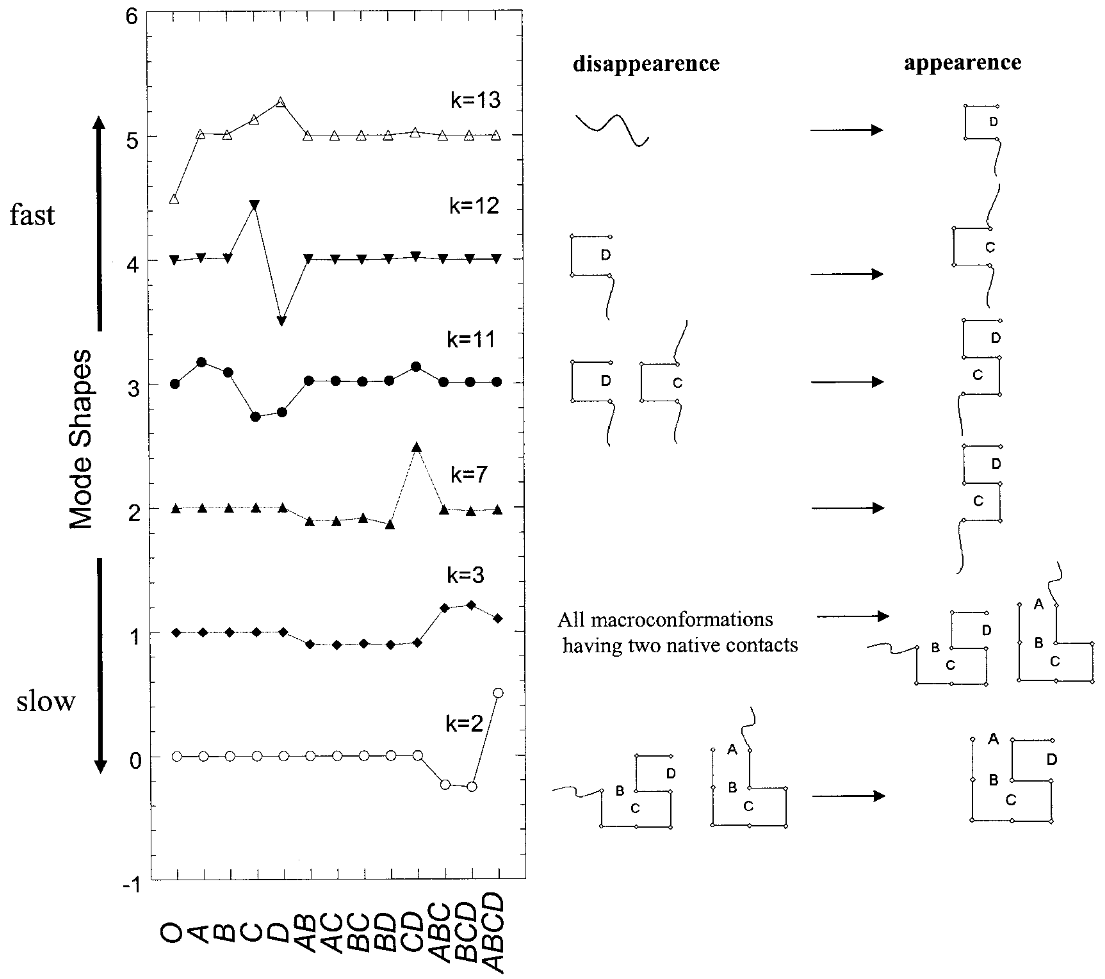


FIGURE 4 Shapes of selected modes k = 2, 4, 7, and 11–13 operating at different stages of the folding of the 9-mer into the native structure *ABCD*. The ordinate displays the eigenvectors (modes) and the abscissa lists the macroconformation ordered as indicated. The extrema on the curves indicate which macroconformations are affected by the particular mode. Note that increasingly more structured macroconformations are activated at lower frequency modes. The macroconformations whose populations increase/decrease by the action of given mode are displayed on the right.

single exponential accumulation of native conformation. At intermediate times, there is a superposition of multiple modes giving rise to a more complex scheme with a multiexponential time dependence.

The decomposition of the reduced transition rate matrix of 16-mer is smoother and broader, as shown in the inset of Figure 3. Approximately five orders of magnitude difference is observed in the time scale of the fast and slow processes. This is consistent with the large-time scale difference observed in two-state folding experiments. The difference is due to two factors: (a) the rate of the individual passages between macrostates and (b) the *multiplicity* of microscopic passages between macrostates characterized by fewer contacts.

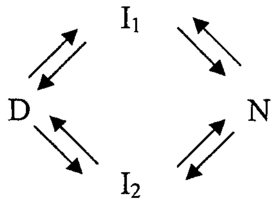
The eigenvalues $\lambda_k{}'$ ($2 \leq k \leq 13$) of **A**′ (13 × 13) for the 9-mer represent the frequencies of the modes in the space of macroconformations, and the eigenvectors describe the transitions driven by that specific mode. There are 12 nonzero eigenvalues. Each element of a given eigenvector is associated with a given macroconformation, the latter being indexed from 1 (macroconformation *0*) to 13 (*ABCD*). The minima or maxima indicate the macroconformations with the highest activity (or transition probability).

The slow and fast modes of the 9-mer in the reduced space of transitions are found to differ by 4–5 orders of magnitude in their frequencies . The dispersion obtained for the 740 × 740 transitions on the other hand, varies over 2–3 orders of magnitude. Comparison of the two sets shows that the slowest modes (k = 2) have about the same frequencies (~0.28/unit time), whereas the fastest modes differ. This difference can be explained as follows: The fast transitions observed in the space of macroconformations reflect the cumulative contribution from the *multiple* transitions, or multiple pathways of relaxation, simultaneously operating at the initial stages of folding, in conformity with the funnel energy landscape view of folding starting from an ensemble of denatured conformations. For example, subset *O* disappears—and subsets *C* and *D* form—by multiple mechanisms, via transitions between several conformations at the initial stage of the folding process, hence the apparent fast transitions (high frequencies) are observed at short times in the space of macroconformations.

Figure 4 illustrates the shapes of a few modes (k = 2, 4, 7, 11–13). These are simply the eigenvectors plotted against macroconformation index. The extrema indicate the macroconformations that are most strongly influenced by the action of a given mode. Positive and negative values refer to changes in opposite direction, i.e., the macroconformation being formed (or accumulated) vs those disrupted (or depleted). The macroconformations whose populations are increased/decreased due to action of given mode are presented in Figure 4. The uppermost curve (k = 13) describes the fastest mode, and the lowermost (k = 2), the slowest. The former, reveals, for example, that the fastest mode decreases the population of subset *O*, while increasing those of subsets *D* and *C*. The second fastest mode (k = 12) describes the communication between subsets *C* and *D*. The third (k = 11) reveals the depletion of *C* and *D*, and concurrent accumulation of *A*, *B*, and *CD*. These three modes lie all in the fast transitions regime (Figure 3). The curve k = 7, on the other hand, reflects an intermediate time process, mainly an equilibration in favor of *CD* between all macroconformations involving two native contacts. Finally, the lowermost two curves refer to the slow increase in the population of conformations having three native contacts (k = 4) and the stabilization of the native structure at the expense of subsets *ABC* and *BCD* (k = 2).

The behavior observed in Figure 4 may be indicative of a general physical relationship between eigenvectors and conformational changes. Accordingly, each individual mode controls one or more specific types of transitions, between particular conformations; and the transitions induced by different modes may be assessed from the maxima and minima of the mode shapes (eigenvectors). To further clarify the physical meaning of the eigenvalues and the eigenvectors of the transition rate matrix, we consider two simple models: series and parallel pathways that involve two intermediates, I_1 and I_2.

$$\mathrm{D} \rightleftarrows \mathrm{I}_1 \rightleftarrows \mathrm{I}_2 \rightleftarrows \mathrm{N}$$



The transition rate matrix that defines the series pathway is

$$\mathbf{A} = \begin{bmatrix} -k_1 & k_{-1} & 0 & 0 \\ k_1 & -(k_{-1}+k_2) & k_{-2} & 0 \\ 0 & k_2 & -(k_{-2}+k_3) & k_{-3} \\ 0 & 0 & k_3 & -k_{-3} \end{bmatrix} \qquad (10)$$

and its counterpart for the parallel pathway is

$$\mathbf{A} = \begin{bmatrix} -(k_1+k_2) & k_{-1} & k_{-2} & 0 \\ k_1 & -(k_{-1}+k_4) & 0 & k_{-4} \\ k_2 & 0 & -(k_{-2}+k_3) & k_{-3} \\ 0 & k_4 & k_3 & -(k_{-3}+k_{-4}) \end{bmatrix} \qquad (11)$$

The decomposition of these matrices gives three nonzero eigenvalues, in each case. Let us adopt the following values for the rate constants for exploratory purposes: Let the consecutive rate constants for the forward reactions in the series model be $k_1 = 0.5$ between D and I_1, $k_2 = 0.1$ between I_1 and I_2, and $k_3 = 0.001$ per unit time between I_2 and N, and the rate constants for the respective reverse reactions k_{-1}= 0.05, $k_{-2} = 0.01$, and $k_{-3} = 0.0001$ per unit time. The rate constants for the forward reactions in the parallel model are taken as $k_1 = 0.5$ between D and I_1, $k_2 = 0.1$ between D and I_2, $k_3 = 0.001$ between I_2 and N, and $k_4 = 0.005$ between I_1 and N, and those of reversible reactions are k_{-1}= 0.05, $k_{-2} = 0.01$, $k_{-3} = 0.0001$, and $k_{-4} = 0.0005$ per unit time. Upon decomposition the transition rate matrix given by Eq. (10), the slowest and fastest modes frequencies are found to be two orders of magnitude different in the series model, whereas those of the parallel model found from the eigenvalue decomposition of Eq. (11) differ by just one order magnitude. If we just consider the forward reactions assuming that the reversible

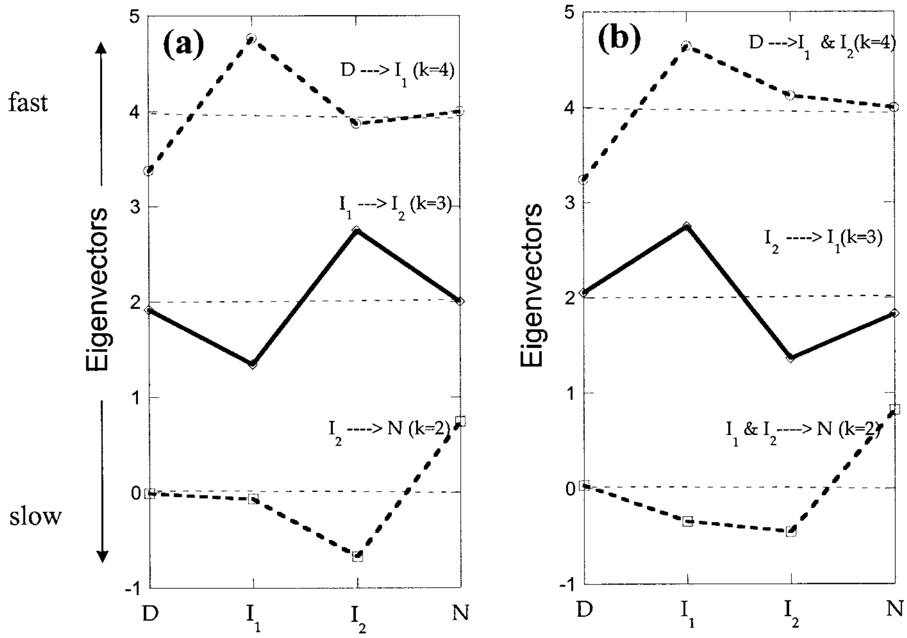


FIGURE 5 Shapes of eigenmodes in descending order for the series (a) and parallel (b) models having two intermediates. The transitions follow a sequential order in the series model, while simultaneous transitions to both intermediates (I_1 and I_2) are observed at the fastest mode in the case of parallel passages. Likewise, the transitions I_1 and I_2 to N occur simultaneously at the slowest mode in part (b).

reactions are negligibly slow, then the eigenvalues of the series model will be k_1, k_2, and k_3, and they will be k_1+k_2, k_3, and k_4 in the parallel model (see Appendix).

Figure 5 presents the plots of the eigenvectors with respect to the states (D, I_1, I_2, and N) for the two models. In the series model [Fig. 5 (a)] we observe the transitions between D to I_1 (uppermost curve, fastest mode, *4*), I_1 to I_2 (middle curve, intermediate mode, *3*) and I_2 to N (lowest curve, slowest mode, *2*). However in the parallel passages model, the transitions from D to I_1 and I_2 concur at the fastest mode. The transition from D to I_1 is relatively more pronounced because of the corresponding higher rate constant ($k_1 = 0.5$ per unit time, as opposed to $k_2 = 0.1$). The curve for the intermediate mode ($k = 3$) shows the accumulation in the population of I_1 and the simultaneous decrease in the population of I_2, basically the competition between two routes, which is actually dominated by the relative rates of depletion of the two intermediate states. Finally, the transitions from I_1 and I_2 to native state occur together at the slowest mode unlike the series model.

The time evolution of each state in the case of a series model (D → I_1 → I_2 → N) were computed using the rate constants ($k_1 = 10^{-2}$, $k_2 = 10^{-4}$, $k_3 = 10^{-6}$) starting from fully unfolded state where at time = 0 , $P_D(0) = 1$ and $P_{I1}(0) = P_{I2}(0) = P_N(0) = 0$. We observe sequential transitions along the pathway [Figure 6(a)]. This series model of sequential intermediates has been referred to as the Staircase Model.[23] The decrease in the population of I_1 precedes the increase in I_2, while the formation of the native state occurs at the expense of state I_2, in conformity with the middle and lower curves in Figure 5(a). On the other hand, the time evolutions of these states exhibit a different behavior for a parallel model (D → I_1 → N and D → I_2 → N) when the computations were performed using the rate constants ($k_1 = 10^{-2}$, $k_2 = 2 \times 10^{-2}$, $k_3 = 10^{-3}$, and $k_4 = 2 \times 10^{-3}$ per unit time) with the same initial conditions as the series model [Fig. 6(b)]. The increase and the decrease in the population of I_1 and I_2 occur at same time interval. However, the change in the population of I_2 is larger than that of I_1 due to the higher rate constant for the transition D to I_2. The formation of the native state starts as I_1 and I_2 are depleted. The detailed formulations for the time evolution of the denatured, intermediate and native states in the case of series and parallel models are presented in the Appendix.

In the case of a broader ensemble of conformations, on the other hand, the time evolution of different macroconformations can obey significantly more complex forms.[24] Figure 7 displays, for example, the time evolution of the macroconformations computed for the 16-mer. The conformations BCDEGHI, ABCDEFG, and ABCGHI all fill up and empty out over the same time course, and hence are not sequential,

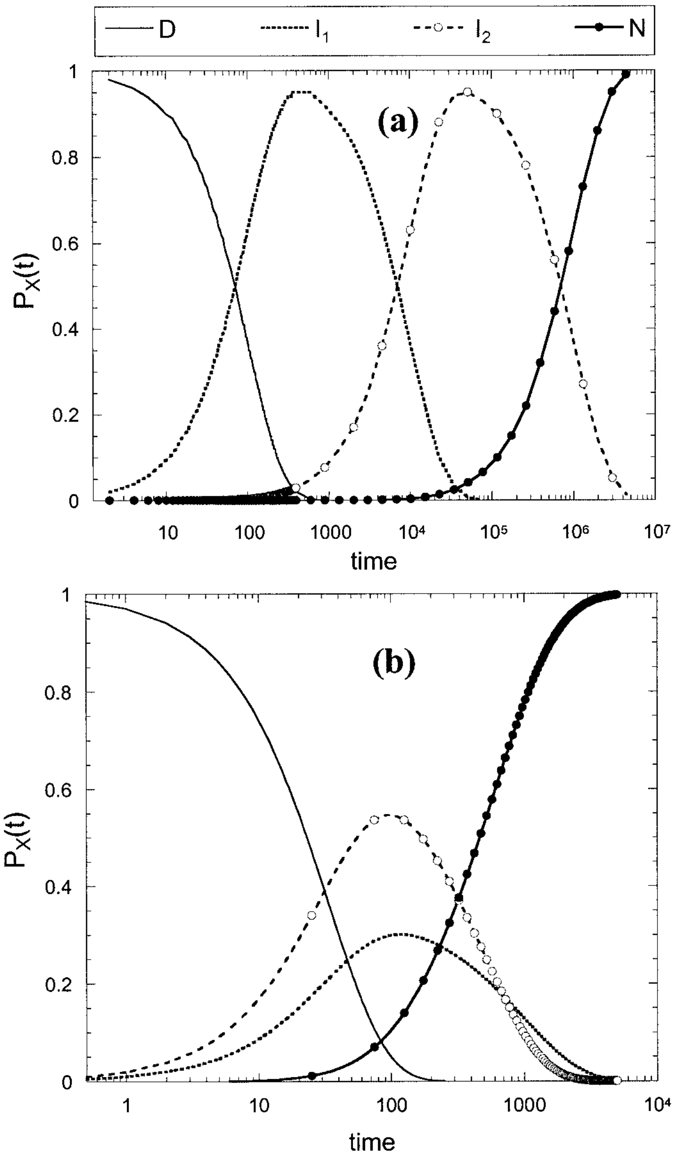


FIGURE 6 Time evolution of each state (a) for the sequential scheme D → I_1 → I_2 → N, using the rate constants 10^{-2}, 10^{-4}, and 10^{-6}/unit time for the respective steps D → I_1, I_1 → I_2 and I_2 → N; and (b) for the parallel scheme D → I_1 → N and D → I_2 → N, using the rate constants 10^{-2}, 2×10^{-2}, 10^{-3}, and 2×10^{-3}/unit time for the respective steps D → I_1, D → $I_{2,}I_2$ → N, and I_1 → N, with the initial conditions $P_U(0) = 1$ and $P_{I1}(0) = P_{I2}(0) = P_N(0) = 0$.

but parallel. Moreover, Figure 7 shows another feature of parallel processes: a slow step is not simply one-contact-more-native than a faster step. There is a fast hidden intermediate state, ABCDEF, which is a simple precursor of ABCDEFG, but it is not a precursor of BCDEGH or ABCGHI.

CONCLUSIONS

We describe the master equation method for analyzing protein folding kinetics, and show how it can be used to identify the intermediate states, and the macrostates that are transiently stabilized as the "transition state," even for nonclassical landscapes such as folding funnels. This method is rigorous and requires no assumptions.

We illustrate the method using a simple Go model that can be analyzed exactly. We find that folding proceeds via a large multiplicity of microscopic routes. The microscopic chain conformations can be conveniently collected into macrostates, resembling those in mass-action models, and classical pathways can be defined in terms of sequences of those macrostates.

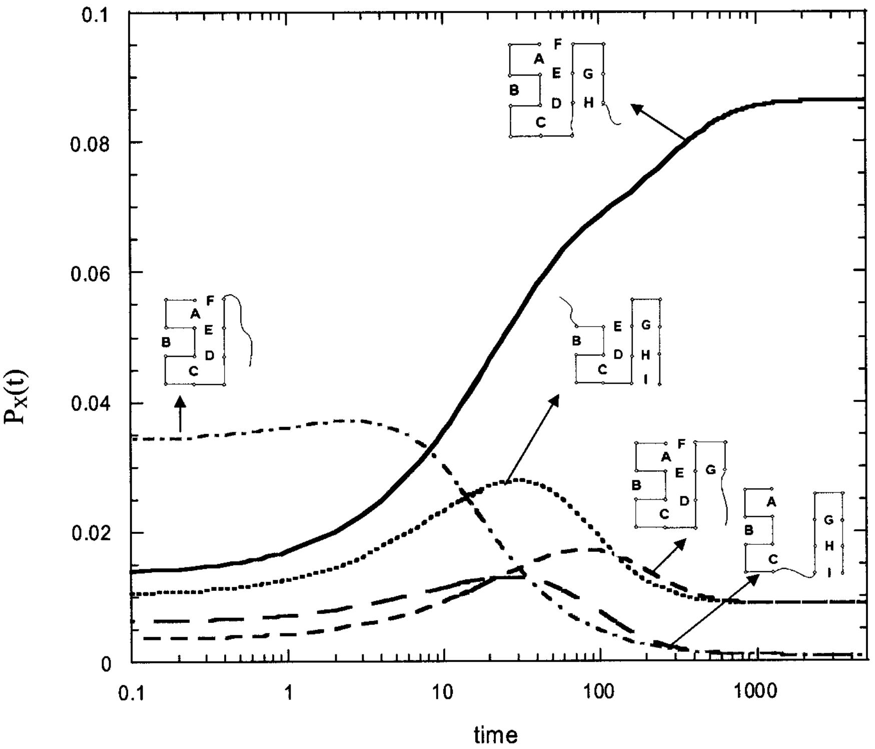


FIGURE 7 Time evolution of substructured 16-mer macroconformations. The peaks indicate the macroconformations that are accumulated before completion of folding. The conformations BC-DEGHI, ABCDEFG, and ABCGHI appear and disappear over the same time course, in a parallel manner.

The transition state can be usually assessed from the shape of the eigenvector corresponding to the slowest nonzero mode(s) of relaxation. This would be a state whose depletion at long times is accompanied by formation of the native state. Furthermore, we should expect this state to be temporarily stabilized (or accumulated) preceding the passage to the slowest mode that completes the folding process. Based on these qualitative features, using the rate constants defined in the text, the respective states I_2 and I_1 appear as the momentarily stabilized transition states, preceding the formation of state N, in the respective cases of the simple parallel and series models with two intermediate states [see Figure 5(a) and (b)]. In the case of more realistic models with multiple conformations [Figure 6(b)], on the other hand, the assessment of the transition state becomes more complicated, because the transition state is a broad ensemble of conformations, not a single well-defined structure. Focusing on the reduced set of macroconformations that are characterized by well-defined subsets of native contacts, appears as a useful conceptual framework for analyzing the dynamics of folding process in this case.

APPENDIX

The set of differential rate equations for each state in the series model of D → I_1 → I_2 → N with rate constants k_1, k_2 and k_3 is

$$d[D]/dt = -k_1[D] \tag{A.1}$$

$$d[I_1]/dt = k_1[D] - k_2[I_1] \tag{A.2}$$

$$d[I_2]/dt = k_2[I_1] - k_3[I_2] \tag{A.3}$$

$$d[N]/dt = k_3[I_2] \tag{A.4}$$

The simultaneous solution of the above set gives the time evolution of each state as

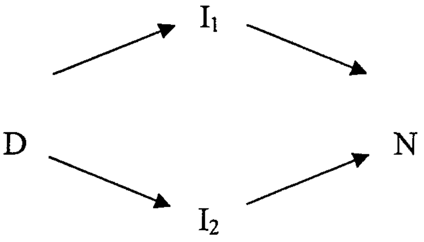
$$[D] = [D]_0 e^{-k_1 t} \tag{A.5}$$

$$[I_1] = [D]_0 \frac{k_1}{k_2 - k_1} [e^{-k_1 t} - e^{-k_2 t}] \tag{A.6}$$

$$[I_2] = [D]_0 \frac{k_1 k_2}{k_2 - k_1} \left[\frac{1}{k_1 - k_3} (e^{-k_3 t} - e^{-k_1 t}) + \frac{1}{k_2 - k_3} (e^{-k_2 t} - e^{-k_3 t}) \right] \quad (A.7)$$

$$[N] = [D]_0 - [D] - [I_1] - [I_2] \quad (A.8)$$

In the case of the parallel model



with the rate constants $k_1(\mathrm{D} \rightarrow \mathrm{I}_1)$, $k_2(\mathrm{D} \rightarrow \mathrm{I}_2)$ $k_3(\mathrm{I}_2 \rightarrow \mathrm{N})$, and $k_4(\mathrm{I}_2 \rightarrow \mathrm{N})$, the set of differential rate equations becomes

$$d[D]/dt = -(k_1 + k_2)[D] \quad (A.9)$$

$$d[I_1]/dt = k_1[D] - k_4[I_1] \quad (A.10)$$

$$d[I_2]/dt = k_2[D] - k_3[I_2] \quad (A.11)$$

$$d[N]/dt = k_4[I_1] + k_3[I_2] \quad (A.12)$$

and the time evolution of the states is obtained from

$$[D] = [D]_0 e^{-(k_1+k_2)t} \quad (A.13)$$

$$[I_1] = [D]_0 \frac{k_1}{k_4 - (k_1 + k_2)} [e^{-(k_1+k_2)t} - e^{-k_4 t}] \quad (A.14)$$

$$[I_2] = [D]_0 \frac{k_2}{k_3 - (k_1 + k_2)} [e^{-(k_1+k_2)t} - e^{-k_3 t}] \quad (A.15)$$

$$[N] = [D]_0 - [D] - [I_1] - [I_2] \quad (A.16)$$

REFERENCES

1. Widom, B. Science 1965, 148, 1555–15560.
2. Widom, B. J Chem Phys 1971, 55, 44–52.
3. Oppenheim, I.; Shuler, K. E.; Weiss, G. H. Adv Mol Relax Processes 1967, 1, 13–68.
4. Bahar, I. J Chem Phys1989, 91, 6525–6531.
5. Gardiner, G. W. Handbook of Stochastic Methods for Physics, Chemistry, and Natural Sciences; Springer-Verlag: London 1990.
6. Van Kampen, N. G. Stochastic Processes in Physics and Chemistry; Elsevier: North Holland, 1992.
7. Leopold, P. E.; Montal;, M.; Onuchic, J. N. Proc Natl Acad Sci USA 1992, 89, 8721–8725.
8. Chan, H. S.; Dill, K. A. J Chem Phys 1993, 99, 2116–2127.
9. Zwanzig, R. Proc Natl Acad Sci USA 1997, 94, 148–150.
10. Zwanzig, R. Proc Natl Acad Sci USA 1995, 92, 9801–9804.
11. Ye, Y. J.; Ripoll, D. R.; Scheraga, H. A. Comp Theor Polymer Sci 1999, 9, 359–370.
12. Cieplak, M.; Henkel, M.; Karbowski, J.; Banavar, J. R. Phys Rev Lett 1998, 80, 3654–3657.
13. Munoz, V.; Henry, E. R.; Hofrichter, J.; Eaton, W. A. Proc Natl Acad Sci USA 1998, 95, 5872–5879.
14. Zhang, W. B.; Chen, S. J. Proc Natl Acad Sci USA 2002, 99, 1931–1936.
15. Ozkan, S. B.; Bahar, I.; Dill, K. A. Nat Struct Biol 2001, 8, 765–769.
16. Stollenwerk, N.; Briggs, K. M. Phys Lett A 2000, 274, 84–91.
17. Goel, N. S.; Richter, Dyn N. Dynamic Stochastic Models in Biology; Academia: New York, 1974.
18. Paci, E.; Vendruscolo, M.; Karplus, M. Proteins 2002, 47, 379–392.
19. Hoang, T. X.; Cieplak, M. J Chem Phys 2000, 113, 8319–8328.
20. Pande, V. S.; Rokhsar, D. S. Proc Natl Acad Sci USA 1999, 96, 1273–1278.
21. Klimov, D. K.; Thirumalai, D. J Mol Biol 1998, 282, 471–492.
22. Jacob, M.; Schmid, F. X. Biochemistry 1999, 38, 13773–13779.
23. Englander, S. W. Ann Rev Biophys Biomol Struct 2000, 29, 213–238.
24. Ozkan, S. B.; Dill, K. A.; Bahar, I. Protein Sci 2002, 11,1958–1970.
25. Ueda, Y.; Taketomi, T.; Go, N. Int J Peptide Res 1975, 7, 445–449.