Recovering Kinetics from a Simplified Protein Folding Model Using Replica Exchange Simulations: A Kinetic Network and Effective Stochastic Dynamics

Weihua Zheng,[†] Michael Andrec,[‡] Emilio Gallicchio,[‡] and Ronald M. Levy*,[‡]

Department of Physics and Astronomy, Rutgers, The State University of New Jersey, 136 Frelinghuysen Road, Piscataway, New Jersey 08854, and Department of Chemistry and Chemical Biology and BioMaPS Institute for Quantitative Biology, Rutgers, The State University of New Jersey, 610 Taylor Road, Piscataway, New Jersey 08854

Received: January 15, 2009; Revised Manuscript Received: July 6, 2009

We present an approach to recover kinetics from a simplified protein folding model at different temperatures using the combined power of replica exchange (RE), a kinetic network, and effective stochastic dynamics. While RE simulations generate a large set of discrete states with the correct thermodynamics, kinetic information is lost due to the random exchange of temperatures. We show how we can recover the kinetics of a 2D continuous potential with an entropic barrier by using RE-generated discrete states as nodes of a kinetic network. By choosing the neighbors and the microscopic rates between the neighbors appropriately, the correct kinetics of the system can be recovered by running a kinetic simulation on the network. We fine-tune the parameters of the network by comparison with the effective drift velocities and diffusion coefficients of the system determined from short-time stochastic trajectories. One of the advantages of the kinetic network model is that the network can be built on a high-dimensional discretized state space, which can consist of multiple paths not consistent with a single reaction coordinate.

1. Introduction

Protein folding is a fundamental problem in modern molecular biophysics and is an example of a slow process occurring via rare events in a high-dimensional configurational space.[1] For this reason, it is difficult for an all-atom simulation to obtain meaningful information on the kinetics and pathways of such processes. A number of strategies for addressing this problem have been proposed over the years that involve focusing on the important slow processes while neglecting the less interesting rapid kinetics by simplification of the state space, reduction of dimensionality, or other methods.[2−21]

If the process in question has an enthalpic or entropic barrier, then most of the time is spent by the system within free-energy basins, while the crossings between basins are relatively rapid but rare. This fact was exploited by Chandler and co-workers in their transition path sampling approach, where a MC procedure is used to sample entire time-ordered paths connecting reactant and product wells in a well-defined manner.[5] While this approach is based on solid statistical mechanical theory and can yield quantitative estimates of the reaction rate, in practice, it remains challenging for large molecular systems with multiple intervening metastable free-energy basins.[22] Another approach to study rare folding events consists of combining information from a large number of short molecular dynamics (MD) trajectories steered by rare events.[6,7,23] In these strategies, multiple replicas are run independently on different processors with different speeds until a transitions occurs in one of the replicas, at which point all replicas are updated to reflect this transition. This coupling of the replicas approximates a single long trajectory with a greatly extended time scale.[23,24] In a similar spirit, the "milestoning" technique makes use of many short simulations spanning predefined critical points along a given reaction path.[8]

A related set of methods for obtaining kinetic information is based on the use of stochastic dynamics on a free-energy landscape.[9−15] They rely on the premise that if one can find a good reaction pathway for the system, then microscopic all-atom dynamics can be used to obtain effective diffusion and drift coefficients along that pathway, allowing the study of the kinetics of the system by low-dimensionality Langevin simulations. While various strategies have been proposed to discover good reaction coordinates in complex systems,[25−27] the fact that the details of the kinetics are projected onto few reaction coordinates can lead to a loss of kinetic information, particularly for systems with complex transition pathways.

Another strategy for improving computational efficiency consists of discretizing the state space and constructing rules for moving among those states. The resulting scheme can be represented as a graph or network,[28] and the kinetics on this graph is often assumed to have Markovian behavior.[16−21] This approach is particularly well suited for reduced lattice models and was first introduced in that context.[16] For systems with a continuous state space, some form of discretization is required. This can be done by clustering based on chosen reduced coordinates[18,28] or using other automated methods.[21,29,30] These clusters must be chosen carefully so as to satisfy the Markovian condition.[19,20,31,32] Alternatively, the discretization can be based on an analysis of the minima and/or saddle points of the energy surface,[17,33,34] which can be used to build a tree-like representation of the potential or free-energy surface (the "disconnectivity graph") or to perform a discretized version of transition path sampling.[35] The location of all minima or saddle points, however, can be a serious challenge for high-dimensional systems, though it has been shown that this is possible for peptide systems.[34,36] A hybrid approach has also been proposed that makes use of molecular dynamics to infer local transition regions to build disconnectivity graphs.[37]

* To whom correspondence should be addressed. E-mail: ronlevy@lutece.rutgers.edu. Phone: 732-445-3947. Fax: 732-445-5958.
[†] Department of Physics and Astronomy.
[‡] Department of Chemistry and Chemical Biology and BioMaPS Institute for Quantitative Biology.

10.1021/jp900445t CCC: $40.75
Published on Web 08/05/2009

While discretization methods based on the clustering of microstates are very powerful, in that they can greatly increase the computational efficiency and allow for the possibility of studying multiple pathways (to the degree that the discretization allows it), they do suffer from some disadvantages. As previously noted,[11,26] a careless choice of reduced coordinate can lead to incorrect kinetics. Furthermore, although a properly constructed kinetic network model will preserve the correct populations of the chosen macrostates, the correctness of populations and potentials of mean force (PMFs) for other reduced coordinates is not guaranteed.

Generalized ensemble methods[38] such as replica exchange molecular dynamics (REMD)[39] have been developed, which enhance the ability to obtain accurate canonical populations in complex systems by increasing sampling efficiency.[40] However, since REMD involves temperature swaps between MD trajectories, it is not straightforward to obtain kinetic information from such simulations.[14,20,41] While approaches based on Markovian models of kinetics between macrostates have been used for this purpose,[32] we have chosen to make use of a kinetic network model[42] in which the nodes correspond to discrete molecular conformations from REMD simulation trajectories (rather than macrostates), and the edges are derived from an ansatz based on structural similarity. While this model was shown to yield physically plausible kinetics,[42] the scheme which was used to weight nodes arising from different simulation temperatures was such that thermodynamic parameters of the system were not exactly preserved.

Here, we present an improved version of that kinetic network model which is guaranteed to reproduce PMFs with respect to any chosen reduced coordinate while allowing the kinetic behavior to be calibrated so as to reproduce the kinetics of the target system. As before, we discretize the multidimensional configurational space of the system by running replica exchange (RE) simulations of the system and collect snapshots which become the nodes of the network. These nodes are then weighted using a scheme based on the temperature-weighted histogram analysis method (T-WHAM),[43] allowing us to obtain correct thermodynamic averages from the RE samples over all simulation temperatures. We then carry out short-time dynamics simulations to derive local drift velocities and diffusion coefficients on suitably chosen reduced coordinates. The network topology and microscopic rate parameters can be adjusted iteratively until agreement is obtained between the drift velocities and diffusion coefficients derived from simulations on the network and those derived from the local dynamics simulations. Since the network is a discretized representation of the system and does not require additional energy and force evaluations, there is a considerable gain in efficiency, allowing us to study much slower kinetic processes than would be accessible using conventional MD. Furthermore, while our local dynamic parameters are estimated on reduced coordinates, the actual kinetic simulation does not occur on those reduced coordinates but rather on the full network. Since the network topology can be constructed based on virtually all degrees of freedom, this allows for multiple pathways and complex transition states. We demonstrate our approach using a folding-like two-dimensional potential and discuss generalizations to the more complex energy landscapes of atomic-level protein simulations.

2. Methods Section

2.1. Kinetics of the Two-Dimensional Potential and the Representation of Drift Velocity and Diffusion Coefficient. We use a two-dimensional potential (Figure 1) described previously,[44] which was constructed to mimic the anti-Arrhenius temperature dependence of the folding rates seen in proteins. This potential was designed to have an energetic barrier when going from the "folded" ($x < 0$) to the "unfolded" ($x \geq 0$) region and an entropic barrier in the reverse direction. The entropic barrier is achieved by imposing a hard wall constraint that limits the space accessible to the folded region. Specifically, the particle can only move in the region $-1 \leq x \leq 1$, $0 \leq y \leq B(x)$, where the boundary function $B(x)$ is a small constant for $x \leq 0$ and an increasing function of x for $x > 0$

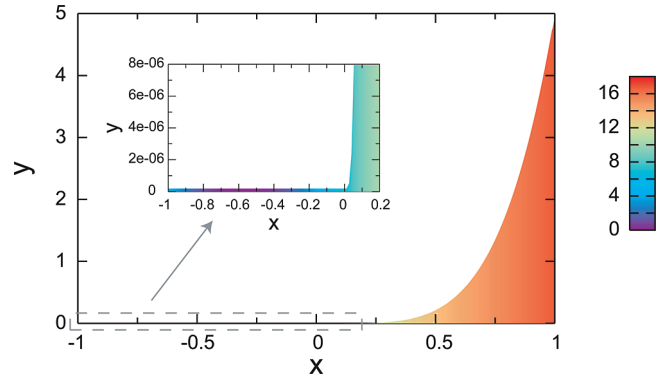

Figure 1. A schematic representation of the two-dimensional potential function used here. The colored area corresponds to the accessible region of the (x,y) plane, with the colors representing the magnitude of the potential energy at that (x,y) point (scale bar in kcal/mol). The potential energy is infinite in the noncolored region and for $y < 0$, $x < -1$, and $x > 1$. The inset is an enlarged view of the folded macrostate and transition region.

$$B(x) = \begin{cases} \delta & 1 \leq x \leq 0 \\ bx^{n_1} + \delta & 0 < x \leq 1 \end{cases} \quad (1)$$

where, for this study, $\delta = 2 \times 10^{-7}$, $b = 5$, and $n_1 = 4.55$. Within this region, the potential energy is given by

$$U(x,y) = \begin{cases} a_1(x + x_0)^2 & -1 \leq x < -x_1 \quad 0 \leq y \leq B(x) \\ -a_2x^2 + c_0 & -x_1 \leq x \leq 0 \quad 0 \leq y \leq B(x) \\ a_3x^2 + c_0 & 0 < x < x_2 \quad 0 \leq y \leq B(x) \\ a_4x^{1/2} + c_1 & x_2 \leq x \leq 1 \quad 0 \leq y \leq B(x) \\ \infty & \text{otherwise} \end{cases} \quad (2)$$

where $a_1 = 23.53$ kcal/mol, $a_2 = 235.3$ kcal/mol, $a_3 = 376.5$ kcal/mol, $a_4 = 11.29$ kcal/mol, and $c_0 = 7.059$ kcal/mol. The dimensionless constants $x_0 = 0.5745$, $x_1 = 0.05222$, and $x_2 = 0.03830$ and the energy offset $c_1 = 5.402$ kcal/mol were chosen so that $U(x,y)$ and its first derivative were continuous. With these parameters, the two-dimensional model has a folding rate with strong anti-Arrhenius temperature dependence, as well as Arrhenius behavior for the unfolding rates in the temperature range studied.

We used Metropolis MC sampling[45] to simulate the movement of a particle in the potential. Because of the large size difference of the accessible region in the y direction between the folded and unfolded regions, we adopted an asymmetric Metropolis−Hastings or "smart MC" proposal scheme.[43,46,47] The step size in the y direction varies with $B(x)$, that is, a proposed move $(\Delta x', \Delta y')$ is generated uniformly in the region of $-\Delta < \Delta x' < \Delta$, $-B(x)\Delta < \Delta y' < B(x)\Delta$, where $\Delta = 0.01$ is a constant for all temperatures. To correct for the asymmetric MC proposal

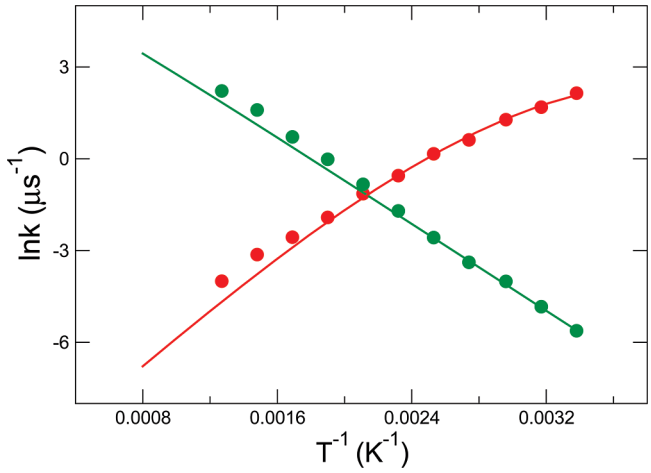

Figure 2. The temperature dependence of the folding and unfolding rate constants. Folding and unfolding rates are indicated by red and green, respectively. The rate constants indicated by circles were derived from kinetic MC simulations run at different temperatures. The lines represent the rates calculated using the Arrhenius equation based on activation energies derived from the PMF along x. Rate constants are expressed in units of 10^{-6} per MC step.

distribution, the Metropolis acceptance probability was multiplied by $\theta(|y' - y|/B(x)\Delta)$ to satisfy detailed balance, where $\theta(z)$ equals 1 if $z < 1$ and 0 otherwise.

Rate constants were obtained via MC simulation by calculating the mean first passage times (MFPTs) in units of MC steps between the two macrostates. A "buffer region" of $-0.1 < x < 0.0437$ was defined as not belonging to either the folded or unfolded macrostate to reduce artifactual rapid recrossings of the barrier. As discussed previously,[44] the folding rate has anti-Arrhenius behavior, that is, it decreases as the temperature increases, as shown in Figure 2. Our goal is to reproduce this temperature dependence of the folding and unfolding rate using a kinetic network model.

The network model is composed of nodes, each representing a microstate of the system, which, in this case, is a point (x,y) in 2D space. We assume that the transition away from a given point along the x coordinate is a diffusive process locally described by a Fokker−Planck equation of the form

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial}{\partial x}[v(x)P(x,t)] + \frac{\partial^2}{\partial x^2}[D(x)P(x,t)] \quad (3)$$

where $P(x,t)$ is the probability density of finding the particle at x at time t and $v(x)$ and $D(x)$ are, respectively, the drift velocity and diffusion constant at x. In this work, $v(x)$ and $D(x)$ are estimated from the rate of change of the mean $\bar{x}(t)$ and variance $\sigma^2(t)$ of the distribution of points obtained from a series of MC trajectories of length t started from points (x,y) at constant x

$$v(x) = \left.\frac{\partial \bar{x}(t)}{\partial t}\right|_{t=t^*} \quad (4)$$

and

$$D(x) = \left.\frac{1}{2}\frac{\partial \sigma^2(t)}{\partial t}\right|_{t=t^*} \quad (5)$$

where t^* is sufficiently large to estimate the quantities of interest but small enough so that it still represents the point of interest x. In practice, $v(x)$ and $D(x)$ are computed by fitting a straight line to $\bar{x}(t)$ and $\sigma^2(t)$ as a function of t. The interval that we used for fitting is five MC steps. It is large enough to include num5 data points and still small enough so that the fit of the parameters is linear to a good approximation. Our goal is to build up a network with kinetics that mimics the local drift velocity and diffusion constant of the MC simulation of the system on the continuous potential.

2.2. Construction of the Kinetic Network Model. ***2.2.1. Discretization of the State Space.*** The nodes of the kinetic network are a discretized approximation of the original state space of the system. We ran a replica exchange Monte Carlo (REMC) simulation on the two-dimensional potential with $S = 8$ replicas at temperatures ranging from 296 to 789 K for 10^9 MC steps. Every 1000 MC steps, transitions between two adjacent temperatures were attempted. Immediately before attempting temperature exchanges, the configuration of each replica was stored, obtaining $N = 50000$ configurations at each temperature and $N \times S = 400000$ configurations at all temperatures. This ensemble of conformations constitutes the discretized state space of the system, which, as described below, approximates well the equilibrium thermodynamics of the system for any temperature within the simulated temperature range or not too far above or below the range.

Traditionally, equilibrium thermodynamic properties of the system at temperature T_0 are obtained by performing canonical sampling at T_0 for a long enough time to obtain convergence. We and others have shown[39,43] that improved convergence can be achieved by employing T-WHAM on RE trajectories over a range of temperatures (which need not include T_0). This yields canonical ensemble averages with greater efficiency than traditional sampling methods because it combines data from high-temperature replicas, which sample high-energy and high-entropy regions, and data from low-temperature replicas, which preferentially sample low-energy, low-entropy regions. The T-WHAM approach is based on a reweighting scheme designed to minimize statistical error.[43] The T-WHAM canonical average $\langle A(T_0)\rangle$ of a quantity A at temperature T_0 is

$$\langle A(T_0)\rangle = \sum_{i=1}^{N} w_i(T_0)A_i \quad (6)$$

where the summation runs over the N RE conformations from all temperatures, A_i is the value of A for conformation i, and the weight factor $w_i(T_0)$ is given by[48,49]

$$w_i(T_0) = \left\{\sum_{k=1}^{S} N_k f_k \exp\left[\left(\frac{1}{k_b T_0} - \frac{1}{k_b T_k}\right)E_i\right]\right\}^{-1} \quad (7)$$

where N_k is the number of samples at each of the S different replica exchange temperatures T_k and k_b is the Boltzmann constant. The natural logarithms of the constants f_k in eq 7 correspond to the relative Helmholtz free energy of each replica k such that $f_k/f_{k'} = Q_{k'}/Q_k$, where Q_k is the canonical partition function of the system at temperature T_k. In T-WHAM, the f_k's are determined by iteratively solving a system of nonlinear equations known as the WHAM equations.[43,48] Thus, each sample i has a weight factor associated with it (eq 7) that depends only on its energy E_i and the temperature of interest T_0. To calculate the PMF of the system as a function of x at

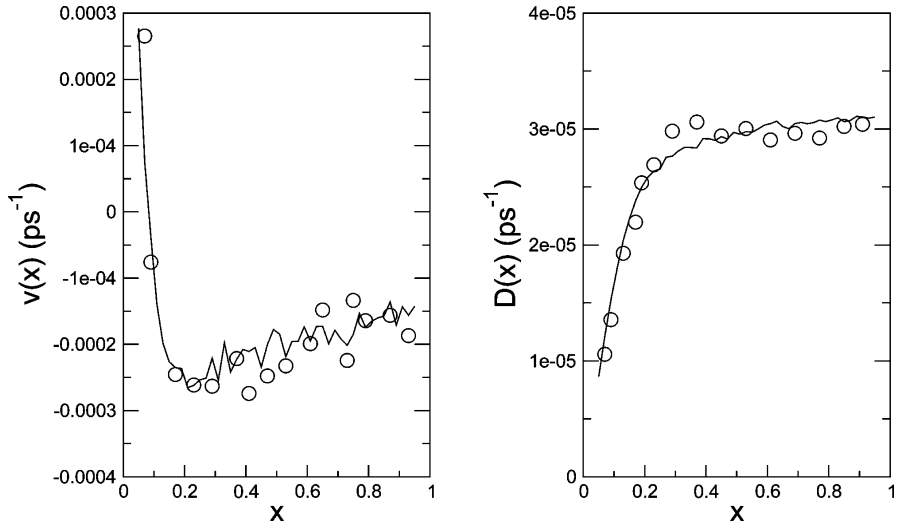

Figure 3. The drift velocity $v(x)$ and diffusion coefficient $D(x)$ along the reaction coordinate x at 298 K. The lines represent the drift velocity and diffusion coefficient of the kinetic MC simulation on the continuous potential, while the circles are the results from the discretized kinetic network model after calibration of c_0.

temperature T_0 using the discretized state space, it is sufficient to employ eq 6 with A being an indicator function which is nonzero if the x-coordinate of the sample is near the designated value of x. This can be done for any temperature T_0, which need not be one of the temperatures used in the RE simulations.

2.2.2. Thermodynamics of the Network Model. To complete the specification of the kinetic network model, we must provide a network topology in the form of edges which connect the nodes and microscopic rates associated with each edge. The choices made for these parameters will determine the kinetics of the network; however, they will not affect the equilibrium thermodynamics of the network as long as detailed balance is satisfied (see eq 8 below) and the network topology is ergodic (i.e., any node is accessible from any other in a finite number of edge traversals). How well the equilibrium properties of the network approximate the real equilibrium thermodynamics of the system depends on the quality of the discretization of the state space using RE.

We connect two nodes with an edge if they are "close" in Euclidean space. Specifically, we join nodes corresponding to coordinates (x,y) and (x',y') if $|x' - x| < \Delta_x$ and $|y' - y| < \Delta_y$. We have chosen the cutoff lengths Δ_x and Δ_y to be much smaller than the dimensions of the system so as to appropriately mimic the local nature of the continuous MC kinetics (see below). We then assign forward and reverse rates to each edge so that detailed balance is satisfied. For example, if nodes i and j are connected by an edge, then we choose rates $k_{ij}(T)$ and $k_{ji}(T)$ such that

$$\frac{k_{ij}(T)}{k_{ji}(T)} = \frac{w_i(T)}{w_j(T)} \tag{8}$$

where $w_i(T)$ and $w_j(T)$ are the weight factors of the two nodes at temperature T, $k_{ij}(T)$ is the rate going from node i to node j, and $k_{ji}(T)$ is the reverse rate. If this detailed balance condition is satisfied, the asymptotic thermodynamics produced by the network model will be the same as that of the original system (subject to the aforementioned ergodicity criterion).

We simulate the kinetics on this network as a continuous time Markov process with discrete states using the Gillespie algorithm.[50] During the simulation, the population histogram along the x coordinate (the reaction coordinate for our two-dimensional system) is accumulated. When a node is visited, its residence time is added to the corresponding bin in the histogram, and at the end of the simulation, the histogram is used to calculate the PMF along the x coordinate.

2.2.3. Calibration of the Kinetic Properties of the Network Model. Although the network design strategy described above guarantees that the correct thermodynamic properties are reproduced, the ability to reproduce the correct kinetics requires additional considerations. Information about the local dynamics of the system in some form is required to obtain a kinetically realistic network. In this section, we illustrate how this can be done for the case of a two-dimensional potential system, where we reproduce the kinetics of a MC simulation on the continuous potential with a network model.

For kinetic MC simulations, the "time" unit is the MC step. The kinetics depends on the move set, which, in our case, was the box defined by the intervals $[-\Delta,\Delta]$ and $[-10B(x)\Delta,10B(x)\Delta]$ for x and y, respectively, and where $\Delta = 0.01$. Note that the magnitude of the allowed moves in the y direction is not constant but depends on x and varies with the size $B(x)$ of the accessible region in the y direction. To recover the kinetics of the MC simulation on the continuous potential, we choose a network topology that mimics the MC move set, as described in the Appendix.

To assign microscopic rates to the edges that satisfy detailed balance, we could choose

$$k_{ij}(T) = \frac{w_j(T)}{w_i(T)}\mu_{ij} \tag{9}$$

and

$$k_{ji}(T) = \mu_{ij} \tag{10}$$

where $\mu_{ij} = \mu_{ji}$ is a base rate to be determined for each pair of nodes i and j to obtain the best agreement with the observed MC kinetics on the continuous potential. To find the appropriate base rates μ_{ij} to match the drift velocity and diffusion coefficients of the network simulation with that of the kinetic

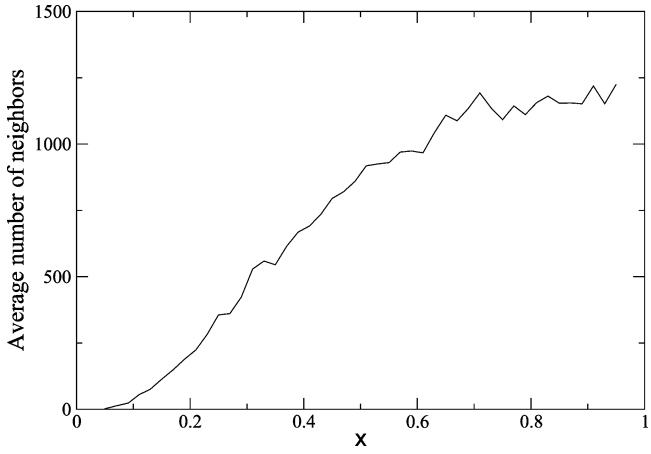

Figure 4. The average number of neighbors per node for all nodes which have a given value of the reaction coordinate x.

MC, we ran 10000 short trajectories (5−10 MC steps) starting at different values of x with both the kinetic MC simulations on the continuous potential and Gillespie simulations on the discretized network model to evaluate the local drift velocities and diffusion coefficients as a function of x. The results are shown in Figure 3.

For the two-dimensional test case studied here, the appropriate values of μ_{ij} are those which allow the network simulation to most closely replicate kinetic MC. In other words, we would like a "time unit" in the Gillespie algorithm to correspond to a MC step in the kinetic MC. In the latter case, each transition between microstates corresponds to an elapsed "time" of 1 unit. Since the edges of the network which join microstates have already been chosen to mimic the kinetic MC move set, it remains only to ensure that the average time between microstate transitions in the discrete network simulation also corresponds to 1 time unit.

In the Gillespie algorithm, the average waiting time in a node is inversely proportional to the sum of the microscopic rates exiting the node. The waiting time in a given node is approximately proportional to the inverse of the number of neighbors for the node, given the fact that the neighbors are close in Euclidean space and have similar energies and similar weights. Therefore, the corresponding edges will have similar rates according to eq 9. As seen in Figure 4, the average number of neighbors for a node increases with x due to the bigger cutoff length in the y direction used to define network edges. Thus, the average waiting time between transitions among microstates will be shorter for nodes with large x. The proportionality between MC steps and Gillespie time units can be maintained by setting $\mu_{ij} = c_0/n_{ij}$, where c_0 is an adjustable coefficient and n_{ij} is the average number of neighbors for the connected nodes i and j. The $1/n_{ij}$ factor in the rate ensures that the waiting times in all nodes are of similar magnitude. We use the average of the number of neighbors for the two connected nodes and not the number of neighbors of the current node since the latter would violate detailed balance if the current and successor nodes have different numbers of neighbors. It should be noted that this strategy for determining μ_{ij} amounts to making the Gillespie algorithm mimic kinetic MC on the continuous potential, and different but analogous approaches would be needed for Newtonian dynamics on a high-dimensional potential.

3. Results and Discussion

To confirm that the 400000 configurations generated using replica exchange MC on the two-dimensional continuous potential give the correct thermodynamic behavior, we compared the PMFs along the x coordinate at several temperatures calculated from the discretized state space and the weight factors of eq 7 with the one calculated by numerical integration of the canonical distribution function of this system. The agreement is excellent at all temperatures examined (only the highest and lowest temperatures are shown in Figure 5 for clarity). This indicates that the correctly weighted discretized state space is a good approximation to the PMF on the continuous potential at all of the temperatures studied. Excellent agreement for the PMF is also obtained from Gillespie simulations using the network model with a generic network topology and rate parameters ($\Delta = 0.01$ and $\mu_{ij} = 1$ for all i,j), as shown in Figure 6. This validates the implementation of the network model algorithm and indicates that the ergodicity condition is satisfied.

We ran a series of short-time trajectories using both kinetic MC on the continuous potential and Gillespie dynamics on the discretized network model and evaluated the drift velocities and diffusion coefficients along the reaction coordinate at different x positions. By varying the parameters of the network in order to match the drift and diffusion on the network with that of the kinetic MC simulation on the conditional potential, we obtained optimized rate parameters for the network model. We made use of the choice of μ_{ij} described above and adjusted c_0 by grid

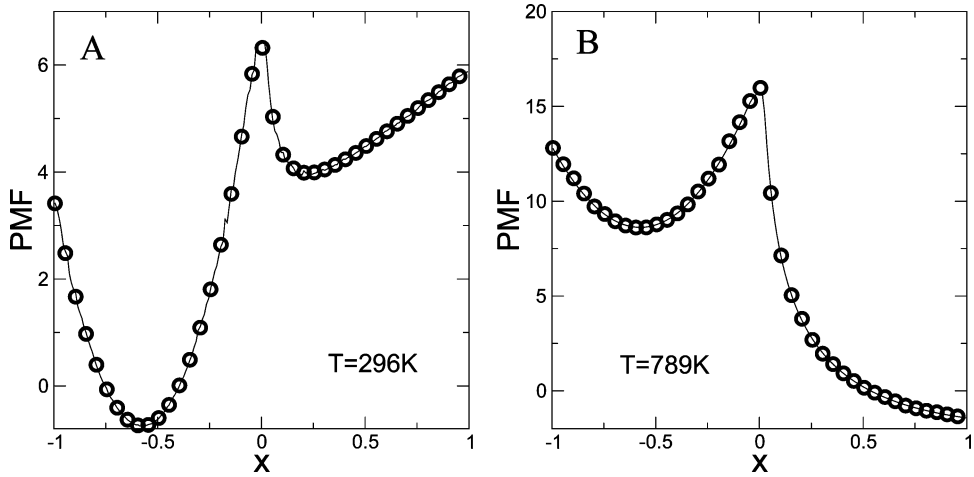

Figure 5. The PMF at two different temperatures, 296 and 789 K. Solid lines are the exact values calculated by numerical integration of the potential. Circles are derived from the full ensemble of eight temperatures combined using WHAM.

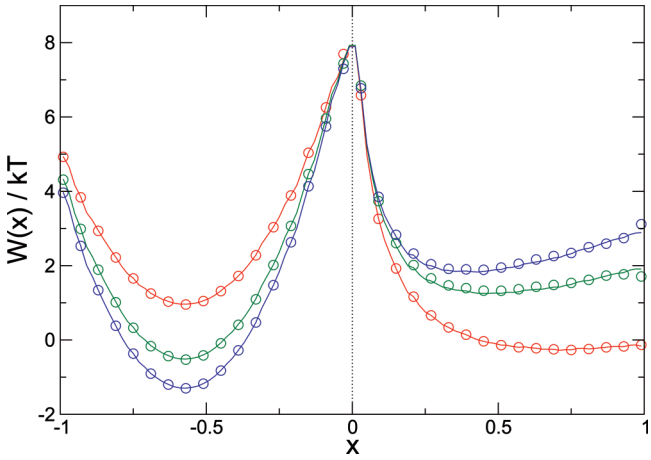

Figure 6. The PMF along the x coordinate at three temperatures, 395, 431, and 526 K (blue, green, and red, respectively). Solid lines are the exact PMFs calculated by numerical integration of the potential, while the circles are derived from kinetic network simulations at each temperature. Approximately 10^3 barrier crossings were observed in the 1 h of wall clock time required for each simulation.

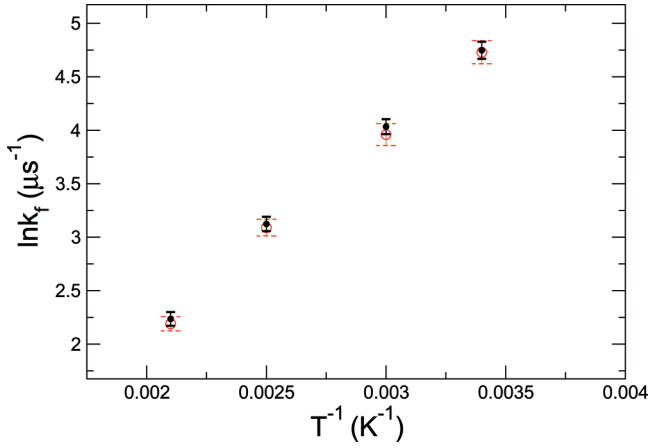

Figure 7. Arrhenius plot of the folding rates of the model system. The solid symbols represent the folding rate from kinetic MC simulations on the continuous potential in units of 10^{-6} per MC step. The open circles represent the rates from simulations of the discretized kinetic network model. The error bars represent one standard deviation, as estimated by error propagation from the first passage time distribution. Solid and dashed error bar lines correspond to the results of kinetic MC and network models, respectively.

search to optimize the agreement of the drift velocities in the kinetic network model and the kinetic MC. The drift velocities were chosen as the target of the optimization since it showed the stronger dependence on c_0. We found that $c_0 = 0.85$ at $T_0 = 298$ K (for all x) gives good agreement, as shown in Figure 3. Furthermore, the folding rates at different temperatures obtained from MC simulations on the continuous potential and from the discretized kinetic network simulation agree very well, as shown in Figure 7.

The results (Figure 7) demonstrate that for the two-dimensional model system for protein folding studied here, it is possible to reconstruct the folding kinetics on a continuous potential using a discrete network model of the type used by Andrec et al.[42] This network model employed an ad hoc method for assigning weights to nodes from different simulation temperatures, while the present model uses weights based on the firm statistical mechanical footing of the T-WHAM method.[43] In fact, the present formulation for the model system studied here reproduces the exact PMFs obtained from numerical integration. This result is general, in that we can expect to obtain PMFs consistent with the RE simulation and T-WHAM without the need for any special choice of reduced coordinate. This is because the f_k factors which appear in eq 7 are directly related to the free energies associated with a given replica. While the WHAM equations themselves require a choice of a reduced coordinate, which one uses to construct the histograms, the resulting f_k factors do not depend on that choice. Correspondingly, although our local dynamic parameters are estimated on a reduced coordinate, the actual kinetic simulation does not occur on that reduced coordinate but rather on the full network, which, by including virtually all degrees of freedom, allows for multiple pathways and transition states.

The model system studied here is sufficiently simple that we can fully confirm the validity of our approach, but it is of course much simpler than any atomic-level molecular model. There is then the question of the applicability of this methodology to such systems. Previous studies[9−14] have shown that it is possible to capture the local kinetics of complex molecular systems using a limited number of degrees of freedom. Concomitantly, we have shown that discrete network models[42] can yield physically plausible pathways for protein folding. Taken together, these observations indicate that the methodology described here will likely be useful to model the kinetics of complex molecular systems.

Nonetheless, the practical implementation of this methodology will require a careful consideration of the additional complexities involved. For example, the large dimensionality of molecular systems may make it difficult to find good reduced coordinates with respect to which local kinetics can be obtained. Since we only require a well-defined measure of local kinetics, we are not limited to methods based on the Fokker−Planck equation (eq 3). For example, one could make use of conformational clustering to define local neighborhoods and then measure within-cluster and between-cluster dynamics using both local dynamics simulations and the kinetic network. The parameters of the network can then be adjusted to optimize the match. Alternatively, local dynamics can be modeled along appropriate generalized coordinates, or the calibration of the network model parameters can be done using kinetic properties that do not depend on a reaction coordinate.

A second layer of complexity that will be involved in application of this methodology to larger systems arises in the adjustment of the network in order to reproduce local dynamics. In the model described above, the choice of network topology (the number of edges and which nodes they connect) was straightforwardly dictated by the move set of the kinetic scheme MC that we were trying to reproduce. Furthermore, because this structure was independent of target temperature, we assumed that the parameter c_0 could be taken to be a constant for all nodes and all temperatures. In a molecular system, these parameters will likely need to be varied, and the determination of the optimal network parameters may require a multidimensional search over topology and rate parameters μ_{ij}.

4. Conclusions

In this paper, we have presented a novel kinetic network strategy for the study of slow time scale processes that extends and improves our previous approach.[42] We used RE simulation and T-WHAM to generate a large set of discrete states with correct thermodynamics. A kinetic network was constructed using the discrete states as nodes using weights that preserve these populations. The microscopic rates between each pair of nodes were calibrated to ensure local dynamics consistent with

those of the full system. The dynamics on the calibrated network was then simulated to study the long time scale kinetics of the system. We have tested the validity of the methods on a two-dimensional continuous potential with anti-Arrhenius kinetics. Application of the kinetic network model to more complex peptide or protein systems will reveal not only high-dimensional kinetic pathways but also will allow us to estimate quantities such as P_{fold}[51] to acquire information about the transition states of the slow process.

Our method differs from previous approaches which decompose the configurational space into a relatively small number of metastable macrostates or project the kinetics onto a chosen reaction coordinate.[13,14,30,52] Our approach does not rely on finding a good reaction coordinate. We compute local stochastic dynamical quantities on a reduced coordinate, but only as a benchmark to calibrate the parameters of a network model constructed from the full discretized state space of the system. The manner in which this calibration is performed can be tailored to the specific demands of the system being studied, and the quantities used for calibration need not be structural coordinates. After calibration, the kinetic simulations are performed on the full network representing a discretization of the high-dimensional state space. This allows for multiple reaction pathways and allows us the flexibility to analyze the dynamics using reduced coordinates of our choosing.

The network model is a Markovian model, like that of other previous approaches,[17−20,32] but instead of using artificially defined macrostates, we use a large number of microstates collected from a RE simulation of the system. This increases the chances of constructing a realistic picture of the kinetics, at the cost of a larger and more complex network. The computational cost of the Gillespie algorithm to perform a single jump between a node and one of its N neighbors scales as $O(\log N)$ in our implementation. From past experience,[42] a much larger network with $\sim 10^6$ states and $>10^9$ edges is still manageable with current computational resources. Since all configurations are precalculated, there is a much lower computational burden than for a comparable all-atom simulation since (for example) potential energies and forces do not need to be evaluated. If necessary, methods for accelerating Gillespie-type simulations that have been developed in the context of chemical reactions and system biology simulations could be used to mitigate the computational burden.[53] We believe that the kinetic network method demonstrated here will be a useful addition to the arsenal of computational methods for the study of slow processes in complex molecular systems.

Acknowledgment. This work was supported in part by THE National Institutes of Health Grant GM 30580.

Appendix

The goal of designing the kinetic network model is to provide the best possible agreement with the kinetic MC simulation on the two-dimensional continuous potential. This goal is more likely to be met if the structure of the network closely mimics the structure of the move set which underlies the kinetic MC. One key choice in the design of the kinetic network is its topology, that is, which pairs of nodes are to be connected by edges. In previous work,[42] we used a simple "box" rule that placed an edge if two nodes were sufficiently close in configuration space. In the case of the MC kinetic scheme used for the two-dimensional potential here, a better choice would more closely mimic the asymmetry of the particular move set used in the MC simulation. In Figure 8, we show the region that a particle starting from a point (x,y) can access and return in two successive MC steps. It consists of the square region excluding the two corners on the left; although the particle could reach the left corners in one step, it is impossible for it to come back to (x,y) in one step. Therefore, in the network model, we also exclude the corresponding node pairs and construct edges only between nodes that satisfy either of the two conditions

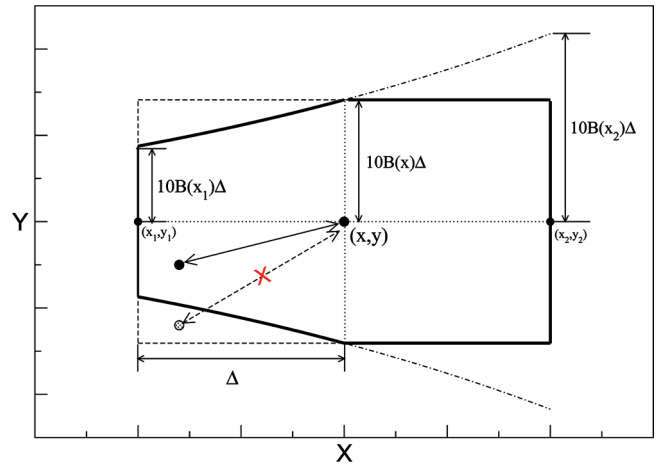

Figure 8. Diagram illustrating the neighboring pair rule for the network model, showing the locus of points (region within the solid line) that can be reversibly visited from a given reference point (x,y). $B(x)$ is the function that defines the accessible region of the system, Δ is the maximum MC step size, and x_1, y_1, x_2, and y_2 correspond to the coordinates of the most distant points reachable from (x,y) in one MC step. The dashed−dotted line encloses the area accessible in one MC move from (x,y). The dashed line is a rectangle around (x,y) of dimensions Δ and $10B(x)\Delta$ along the x and y axes, respectively.

$$\begin{array}{lll} x - \Delta < x' < x & \text{and} & |y - y'| < 10B(x')\Delta \\ x < x' < x + \Delta & \text{and} & |y' - y| < 10B(x)\Delta \end{array} \qquad (11)$$

References and Notes

(1) Onuchic, J. N.; Wolynes, P. G. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70–75.
(2) Elber, R. *Curr. Opin. Struct. Biol.* **2005**, *15*, 151–156.
(3) Snow, C. D.; Sorin, E. J.; Rhee, Y. M.; Pande, V. S. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 43–69.
(4) Christen, M.; van Gunsteren, W. F. *J. Comput. Chem.* **2007**, *29*, 157–166.
(5) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.
(6) Zagrovic, B.; Sorin, E. J.; Pande, V. *J. Mol. Biol.* **2001**, *313*, 151–169.
(7) Pande, V. S.; Baker, I.; Chapman, J.; Elmer, S. P.; Khaliq, S.; Larson, S. M.; Rhee, Y. M.; Shirts, M. R.; Snow, C. D.; Sorin, E. J.; Zagrovic, B. *Biopolymers* **2003**, *68*, 91–109.
(8) Faradjian, A. K.; Elber, R. *J. Chem. Phys.* **2004**, *120*, 10880.
(9) Schumaker, M. F.; Pomès, R.; Roux, B. *Biophys. J.* **2000**, *79*, 2840–2857.
(10) Hummer, G.; Kevrekidis, I. G. *J. Chem. Phys.* **2003**, *118*, 10762.
(11) Kopelevich, D. I.; Panagiotopoulos, A. Z.; Kevrekidis, I. G. *J. Chem. Phys.* **2005**, *122*, 044908.
(12) Best, R. B.; Hummer, G. *Phys. Rev. Lett.* **2006**, *96*, 228104.
(13) Yang, S.; Onuchic, J. N.; Levine, H. *J. Chem. Phys.* **2006**, *125*, 054910.
(14) Yang, S.; Onuchic, J. N.; Garcìa, A. E.; Levine, H. *J. Mol. Biol.* **2007**, *372*, 756–763.
(15) Berezhkovskii, A.; Szabo, A. *J. Chem. Phys.* **2005**, *122*, 014503.
(16) Ozkan, S. B.; Dill, K. A.; Bahar, I. *Protein Sci.* **2002**, *11*, 1958–1970.
(17) Ye, Y.-J.; Ripoll, D. R.; Scheraga, H. A. *Comput. Theor. Polym. Sci.* **1999**, *9*, 359–370.
(18) Singhal, N.; Snow, C. D.; Pande, V. S. *J. Chem. Phys.* **2004**, *121*, 415–425.
(19) Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571–6581.

(20) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward, T. J. C.; Zhestkov, Y.; Zhou, R. *J. Phys. Chem. B* **2004**, *108*, 6582–6594.
(21) Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. *J. Chem. Phys.* **2007**, *126*, 155102.
(22) Rogal, J.; Bolhuis, P. G. *J. Chem. Phys.* **2008**, *129*, 224107.
(23) Voter, A. F. *Phys. Rev. B* **1998**, *57*, R13985–R13988.
(24) Shirts, M. R.; Pande, V. S. *Phys. Rev. Lett.* **2001**, *86*, 4983–4987.
(25) Ma, A.; Dinner, A. R. *J. Phys. Chem. B* **2005**, *109*, 6769–6779.
(26) Krivov, S. V.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 12689–12698.
(27) Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. *J. Phys. Chem. B* **2008**, *112*, 8701–8714.
(28) Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299–306.
(29) Schultheis, V.; Hirschberger, T.; Carstens, H.; Tavan, P. *J. Chem. Theory Comput.* **2005**, *1*, 515–526.
(30) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. *J. Chem. Phys.* **2007**, *126*, 155101.
(31) Chekmarev, D. S.; Ishida, T.; Levy, R. M. *J. Phys. Chem. B* **2004**, *108*, 19487–19495.
(32) Buchete, N.-V.; Hummer, G. *Phys. Rev. E* **2008**, *77*, 030902.
(33) Wei, G.; Mousseau, N.; Derreumaux, P. *Proteins* **2004**, *56*, 464–474.
(34) Evans, D. A.; Wales, D. J. *J. Chem. Phys.* **2004**, *121*, 1080–1090.
(35) Wales, D. J. *Mol. Phys.* **2002**, *100*, 3285–3305.
(36) Carr, J. M.; Wales, D. J. *J. Phys. Chem. B* **2008**, *112*, 8760–8769.
(37) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766–14770.
(38) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96–123.
(39) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
(40) Zheng, W.; Andrec, M.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 15340–15345.
(41) van der Spoel, D.; Seibert, M. M. *Phys. Rev. Lett.* **2006**, *96*, 238102.
(42) Andrec, M.; Felts, A. K.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6801–6806.
(43) Gallicchio, E.; Andrec, M.; Felts, A. K.; Levy, R. M. *J. Phys. Chem. B* **2005**, *109*, 6722–6731.
(44) Zheng, W.; Andrec, M.; Gallicchio, E.; Levy, R. M. *J. Phys. Chem. B* **2008**, *112*, 6083–6093.
(45) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1091.
(46) Hastings, W. K. *Biometrika* **1970**, *57*, 97–109.
(47) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: Oxford, U.K., 1987.
(48) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.
(49) Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *129*, 124105.
(50) Gillespie, D. T. *Markov Processes: An Introduction for Physical Scientists*; Academic Press: Boston, MA, 1992.
(51) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. *J. Chem. Phys.* **1998**, *108*, 334–350.
(52) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Dill, K. A. *Multiscale Model. Simul.* **2006**, *5*, 1214–1226.
(53) Gillespie, D. T. *Annu. Rev. Phys. Chem.* **2007**, *58*, 35–55.

JP900445T