SIMPLIFIED MODELS FOR SIMULATING REPLICA EXCHANGE SIMULATIONS AND RECOVERING KINETICS OF PROTEIN FOLDING

BY WEIHUA ZHENG

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Physics and Astronomy

Written under the direction of

Dr. Ronald M. Levy

and approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2009

ABSTRACT OF THE DISSERTATION

Simplified Models for simulating replica exchange simulations and recovering kinetics of protein folding

by Weihua Zheng

Dissertation Director: Dr. Ronald M. Levy

Protein folding is a fundamental problem in modern structural biology. The nature of the problem poses challenges to the understanding of the process via computer simulations. One of the challenges in the computer simulation of proteins at the atomic level is the efficiency of sampling conformational space. Replica exchange (RE) methods are widely employed to alleviate the difficulty. To study how to best employ RE to protein folding and binding problems, We constructed a kinetic network model for RE studies of protein folding and used this simplified model to carry out ”simulations of simulations” to analyze how the underlying temperature dependence of the conformational kinetics and the basic parameters of RE all interact to affect the number of folding transitions observed. When protein folding follows anti-Arrhenius kinetics, we observe a speed limit for the number of folding transitions observed at the low temperature of interest, which depends on the maximum of the harmonic mean of the folding and unfolding transition rates at high temperature. The efficiency of temperature RE was also studied on a more complicated and realistic continuous two-dimensional potential. Comparison of the efficiencies obtained

using the continuous and discrete models makes it possible to identify non-Markovian effects which slow down equilibration of the RE ensemble on the more complex continuous potential. In particular, the efficiency of RE is limited by the timescale of conformational relaxation within free energy basins. The other challenges we are facing in all-atom simulations is to obtain meaningful information on the slow kinetics and pathways of folding. We present a kinetic network model which recover the kinetics using RE-generated states as the nodes of a kinetic network. Choosing the appropriate neighbors and the microscopic rates between the neighbors, the correct kinetics of the system can be recovered by running a simulation on the network.

Acknowledgements

First, I would like to express my deep gratitude to my adviser Prof. Ronald M. Levy for his passionate inspiration and insightful guidance.

I would also like to thank Dr. Michael Andrec and Dr. Emilio Gallicchio. We had such a wonderful time working shoulder to shoulder on various projects. I've learned not only countless hands-on skills on practical issues but also the commitment and work ethic.

My sincere appreciation goes to the members of Levy's group, graduate students Omar Haq, Chitra Narayanan, Kristina Paris, Mauro Lapelosa and Dr. Daniel Weinstock, Dr. Anthony Felts, Dr. Hisashi Okumura, you all make the lab a warm and enlightening place to work.

Thanks for the great administration work done by Janice Pawlo, Katherine Lam, Jennifer Schenk, Kevin Abbey, Bill Abbott and Peter Talley, it makes my life so much easier than it could be.

Thanks for my friends Xinjie Wang, Chenglin Zhang, Guohui Zheng, Tian Feng, Fei Xu, Bin Zhang and Bin Miao.

Last but not least, I want to thank my wife and my parents for their love and support which help me pull through the hard time.

Dedication

To my dear wife Juan Zhang who gives me the courage and hope in finishing my thesis.

Table of Contents

Abstract ii

Acknowledgements iv

Dedication v

List of Tables ix

List of Figures x

1. Introduction—-Two challenges posed to computer simulations of protein folding 1

2. Simulating Replica Exchange Simulations of protein folding with a kinetic network model 7

2.1. Introduction to Replica Exchange Method 8

2.2. Construction of the network model 10

2.2.1. Thermodynamic model for anti-Arrhenius behavior 15

2.3. Results and Discussion 18

2.3.1. Convergence efficiency of non-RE simulations 18

2.3.2. Convergence efficiency of the kinetic network model for large α limit 20

2.3.3. Convergence efficiency of the kinetic network for finite α 26

2.3.4. Convergence efficiency of the kinetic network under special conditions 28

2.4. Conclusions 32

2.5. Appendix I: Closed form analysis for the $\alpha \to \infty$ limit of the network model 36

2.6. Appendix II: Publication attached 40

3. Simulating Replica Exchange simulation of Protein Folding with a continuous two-dimensional potential model 42

3.1. Introduction 42

3.2. Methods 45

3.2.1. The two-dimensional continuous potential 45

3.2.2. Kinetics on the two-dimensional continuous potential 47

3.2.3. RE simulation on the two-dimensional continuous potential 48

3.2.4. Review of the discrete Network Replica Exchange (NRE) 49

3.3. Results and Discussion 52

3.3.1. Thermodynamics and kinetics of the continuous model system . . . 52

a. Thermodynamics 52

b. Kinetics 56

3.3.2. RE simulations using MC on the continuous potential 62

3.3.3. Non-Markovian effects revealed by comparison of continuous and discrete RE simulations 64

3.3.4. Dependence of RE efficiency on the number of replicas 71

3.4. Conclusions 75

3.5. Appendix I —-The alternative potential function 76

3.6. Appendix II: Publication attached 77

4. Recovering Folding Kinetics From Replica Exchange Simulations With a Kinetic Network Calibrated Using Local Dynamics 79

4.1. Introduction 79

4.2. Methods 82

4.2.1. Kinetics of the two-dimensional potential and the representation of drift velocity and diffusion coefficient 82
4.2.2. Discretization of the state space 86
4.2.3. Thermodynamics of the network model 88
4.2.4. Calibration of the kinetic properties of the network model 89
4.3. Results and Discussion 93
4.4. Conclusions 96
4.5. Appendix 99
References 100
Vita 109

List of Tables

3.1. Number of temperature-conditional transition events in 2×10^9 MC steps for two replicas (with temperatures of 296 K and 474 K) as a function of the number of MC steps between attempted temperature exchanges (N_X), and observed temperature-conditional mean first passage times (in units of 10^6 MC steps, see text for details). 66

3.2. Empirical "reverse-engineered" rates at temperatures $T_1 = 296$ K and $T_2 = 474$ K (in units of 10^{-6} MC step) from continuous potential simulation data assuming the network topology of Figure 3.2 66

3.3. History dependent and independent branching probabilities from state U_2F_1. . . . 78

List of Figures

2.1. Rough energy landscapes of protein folding trap simulations in a local minimum when conventional sampling methods are used(e.g. Monte Carlo or Molecular Dynamics). Using Replica Exchange method, the low temperature replica borrows fast kinetics from high temperature replica to help escaping the local minimum at low temperature. 11

2.2. The kinetic network of the composite states corresponding to the simplified replica exchange model with two replicas. The state labels represent the conformation (letter) and temperature (subscript) for each replica. For example, F_2U_1 represents the state in which replica 1 is folded and at temperature T_2, while replica 2 is unfolded and at temperature T_1. Red and black arrows correspond to folding and unfolding transitions, respectively, while the temperature at which the transition occurs is indicated by the solid and dashed lines (for T_2 and T_1, respectively). The cyan arrows correspond to temperature exchange transitions, with the solid and dashed lines denoting transitions with rate parameters α and $w\alpha$, respectively. . . 13

2.3. Arrhenius plot of the folding and unfolding rates from a thermodynamic model for the temperature dependence of protein folding rate constants. Black line corresponds to unfolding rate, while red lines correspond to the folding rates. The solid line is for the $\Delta C_p^\dagger \neq 0$ case displaying anti-Arrhenius behavior, while the dashed line corresponds to the same parameters with $\Delta C_p^\dagger = 0$. The arrow indicates the temperature T^* at which the folding rate is maximal (≈ 380 K). 17

2.4. Estimates of the relative population of the F conformation at temperature $T_1 = 300$ K for a finite simulation time. The temperature of replica 1 was held fixed at 300 K, while T_2 (of replica 2) is swept from 300 K to 700 K. The temperature exchange parameter α was set to 2 ns^{-1}. For each individual T_2, the system was simulated for $\tau = 1.25$ μs beginning in the state F_1F_2 at time $t = 0$ and the fraction folded at T_1 $S_1(\tau)$ was calculated. This was repeated 50,000 times, and the resulting $S_1(\tau)$ values were averaged and the results are plotted. The solid line corresponds to the anti-Arrhenius folding rates ($\Delta C_p^\dagger \neq 0$), while the dashed line corresponds to the Arrhenius rates ($\Delta C_p^\dagger = 0$) (Figure 2.3). The true fraction folded at $T_1 = 300$ K is the same for both the Arrhenius and anti-Arrhenius models and is indicated by the dotted line. The temperature at which the bias is minimized for the anti-Arrhenius model (≈ 440 K) is indicated by the arrow. 21

2.5. Number of transition events in NRE simulations (normalized by the number of replicas) for various temperature ranges, exchange rates α, and number of replicas N. In all cases, the system was simulated for $\tau = 4$ μs. For the simulations in (A), α was set to 1000 μs^{-1}, the dashed and solid lines correspond to Arrhenius and anti-Arrhenius kinetics, respectively, and six replicas were exponentially distributed between 300 K and T_{max}. The simulations in (B) were performed with anti-Arrhenius rates, N replicas exponentially distributed from 300 K to 700 K, and α values of 10000 μs^{-1} (black), 1000 μs^{-1} (red), 100 μs^{-1}(green), 10 μs^{-1}(blue) and 1 μs^{-1}(cyan). 23

2.6. Number of transition events per replica in NRE simulations using the anti-Arrhenius folding rates for a simulation time $\tau = 4$ ms conditional on temperature $T_1 = 300$ K, while T_2 is scanned from 300 K to 700 K. (A) Black and green solid lines: simulation results for two-replica and three replica systems (with $T_3 = 440$ K), respectively. Black and green dashed lines: number of transition events predicted using the average of harmonic means for two and three replicas, respectively. All simulations were performed with $\alpha = 10$ ns^{-1}. (B) Results for two-replica NRE simulations using the anti-Arrhenius folding rates and α values of 10 ns^{-1} (black solid), 1 ns^{-1} (red), 100 μs^{-1} (green), 10 μs^{-1} (blue), and 1 μs^{-1} (cyan). The black dashed line corresponds to the predicted number of transitions for a single, uncoupled simulation at T_1. 25

2.7. Number of transition events per replica as a function of α for a temperature-independent rate system in a total simulation time of 4 ms. The folding and unfolding rates were those of the anti-Arrhenius model at 440 K (i.e. $k_u = 12.06$ μs^{-1} and $k_f = 1.052$ μs^{-1}). The predicted number of transition events for an uncoupled, non-RE simulation with the same rates and simulation time is shown as a black dashed line and corresponds to the $\alpha \to \infty$ limit. The black, red and green data correspond to $N = 2$, 10, and 40, respectively. 30

2.8. Number of transition events per replica as a function of the number of replicas N for a temperature-independent rate system in a total simulation time of 4 ms. The folding and unfolding rates were those of the anti-Arrhenius model at 440 K (i.e. $k_u = 12.06$ μs^{-1} and $k_f = 1.052$ μs^{-1}). The curves correspond to $\alpha = 5$, 10, 20, 40, 60, 100, and 200 μs^{-1} from bottom to top, respectively. 31

2.9. The average amount of time t spent in a given excursion in temperature space away from the temperature of interest for a temperature-independent rate system as a function of the number of replicas N. The folding and unfolding rates were those of the anti-Arrhenius model at 440 K (i.e. $k_u = 12.06\ \mu\text{s}^{-1}$ and $k_f = 1.052\ \mu\text{s}^{-1}$) and $\alpha = 100\ \mu\text{s}^{-1}$. The line is a least-squares fit. 33

2.10. The collapsed 4-state kinetic network model. 41

2.11. The collapsed kinetic network model with an absorbing state corresponding to walker 1 unfolded. 41

3.1. A schematic representation of the two-dimensional potential function used in this work. 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 non-colored region and for $y < 0$, $x < -1$, and $x > 1$. The inset is an enlarged view of the folded macrostate and transition region. 46

3.2. The kinetic network model for the discrete NRE model used in chapter 2. The state labels represent the conformation (letter) and temperature (subscript) for each replica. For example, F_2U_1 represents the state in which replica 1 is folded and at temperature T_2, while replica 2 is unfolded and at temperature T_1. Red and black arrows correspond to folding and unfolding transitions, respectively, while the temperature at which the transition occurs is indicated by the solid and dashed lines (for T_2 and T_1, respectively). The cyan arrows correspond to temperature exchange transitions, with the solid and dashed lines denoting transitions with rate parameters α and $w\alpha$, respectively. 50

3.3. The potential of mean force (PMF) at three different temperatures: 296 K (solid), 474 K (dashed) and 789 K (dotted). The PMF was calculated using numerical integration. To more clearly illustrate the change in the barrier height as a function of temperature, the three curves have been superimposed to coincide at $x = 0$. . . 53

3.4. The temperature dependence of the fractional population folded (solid line) calculated by numerical integration of the potential of mean force. The temperature dependence of the fraction folded corresponding to a system with a smaller average potential energy difference between the folded and unfolded states (see Appendix I) is shown for comparison (dashed line). The fraction folded derived from the folding and unfolding rates obtained by MC simulation (Figure 3.6) are shown as circles. The melting temperature $T_M = 463$ K (corresponding to 50% folded population) is indicated. 55

3.5. The distributions of first passage times for folding (black) and unfolding (red) observed during a 2.7×10^{10}-step kinetic MC at 475 K. Approximately 4700 folding and unfolding events were observed. A folding first passage time is defined as the time elapsed from when the particle enters the unfolded region from the buffer region (having previously been in the folded region), until it re-enters the folded region. The unfolding first passage time is defined similarly. The semi-log plot of the histograms of the first passage times is shown as circles, while the lines represent the best-fit exponential curve. 57

3.6. The temperature dependence of the folding and unfolding rate constants (solid lines and symbols). Folding and unfolding rates are indicated by red and green color, respectively. The folding and unfolding rates corresponding to a system with a smaller activation energy for folding (Appendix I) are shown for comparison (dashed lines). The rate constants plotted in symbols were derived from kinetic MC simulations run at different temperatures. The solid lines represent the rates calculated using the Arrhenius equation based on activation energies derived from the PMF along x (Figure 3.3). Rate constants are expressed in units of 10^{-6} per MC step. 58

3.7. Number of direct round trip transition events N_{direct} in single temperature uncoupled simulations over the temperature range $296 - -789\ K$ in 5×10^9 MC steps. The curve plotted as a solid line was calculated from the harmonic mean of the folding and unfolding rates estimated from the mean of the folding and unfolding first passage time distribution (Figure 3.5) obtained by MC simulations at each temperature, while the number of events counted directly from the MC simulations at individual temperatures are plotted as circles. The high level of agreement indicates that the system is very well approximated as a two-state activated process. 61

3.8. The dependence of the number of temperature-conditional transition events N_{TE} (section 3.2.3) on the temperature of the high-temperature replica for a two-replica simulation on the continuous potential (circles), and comparison with predicted transition events using discrete NRE model (Figure 3.2) (curves). For all simulations, one replica was fixed at 296 K and the other replica was scanned from 296 K to 789 K. The black dashed line corresponds to the discrete model prediction in the large-α limit. The solid curves are the predicted N_{TE} using the NRE model with temperature dependent folding and unfolding rates taken from the kinetic MC simulations on the continuous potential (shown in Figure 3.6) and using an α corresponding to an attempted temperature exchange. The circles are the N_{TE} values observed in 2×10^9 MC step RE simulations on the continuous potential. The green, red, and blue data correspond to N_X values of 1 000, 200, and 20, respectively. 63

3.9. Number of transition events N_{TE} (section 3.2.3) normalized by the number of replicas in 2×10^9 MC steps for 2 to 15 replicas exponentially distributed in temperature from 296 to 789 K. Temperature exchanges were attempted every 10 000 (solid), 1 000 (dashed), and 200 (dotted) MC steps. 72

3.10. Number of transition events N_{TE} (section 3.2.3) observed in 2×10^9 MC steps for three different 11-replica RE simulations performed using the continuous potential with $N_X = 200$. The temperature distributions for the three simulations are shown in the inset. Transition counts have been normalized by the N_{TE} of simulation A. 74

4.1. A schematic representation of the two-dimensional potential function used in previous chapter. 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 non-colored region and for $y < 0$, $x < -1$, and $x > 1$. The inset is an enlarged view of the folded macrostate and transition region. 83

4.2. 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 simulation 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. 87

4.3. The PMF at two different temperature 296 K and 789 K. Solid lines are the exact value calculated by numerical integration of the potential. Circles are derived from the full ensemble of 8 temperatures combined using WHAM. 90

4.4. The PMF along the x coordinate at the three temperatures 395 K, 431 K, 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. 90

4.5. 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, x_1, y_1, x_2, 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, ... 91

4.6. 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, while the circles are the results from the kinetic network model after calibration of c_0. 97

4.7. The average number of neighbors per node for all nodes which have a given value of the reaction coordinate x. 97

4.8. Arrhenius plot of the folding rates of the model system. The line represents the folding rate from kinetic MC simulation in unit of 10^{-6} per MC step. The circles represent the rates from simulation of the kinetic network model. 98

Chapter 1

Introduction—-Two challenges posed to computer simulations of protein folding

Protein folding is a fundamental problem in modern structural biology, and is an example of a slow process occurring via rare events in a high-dimensional configuration space[1]. The nature of the problem poses two major challenges to the understanding of the folding process via computer simulations. One of the challenges in the computer simulation of proteins at the atomic level is the efficiency of sampling conformational space. The efficiency of many common sampling protocols, such as Monte Carlo (MC) and molecular dynamics (MD) is limited by the need to cross high free-energy barriers between conformational states and rugged energy landscapes. One general class of methods for overcoming this problem involves the use of generalized ensembles[2] which distorts the energy landscape in a way that allows for increased efficiency but which can be ”undone” by appropriate reweighting to recover the canonical ensemble. The most well-known of these approaches is umbrella sampling[3], in which biasing potentials are used to allow for more efficient sampling in regions of high free energy connecting minima of interest. A series of simulations with a set of biasing potentials spanning the reaction coordinate of interest can then be combined using the WHAM method to obtain a potential of mean force along that coordinate[4]. Umbrella sampling has been used extensively in many areas of computational chemistry and physics, including the study of folding [5] and allosteric transitions[6] in proteins. Multicanonical simulation[7, 8] is another generalized ensemble method that can be viewed as an extreme form of umbrella sampling[9], in which a biasing potential

is added in order to make the resulting energy distribution uniform. This allows the system to undergo free diffusive motion in energy space, allowing barriers to be surmounted. Alternatively, the temperature could be made a dynamical variable and a biasing potential could be applied to make the temperature distribution uniform, leading to the simulated tempering algorithm[10]. All of these methods require substantial prior knowledge about the system being studied: a good choice of reaction coordinate must be determined or an appropriate biasing function must be found (often at significant computational cost).

Another class of methods for studying equilibrium properties of quasi-ergodic systems that has received a great deal of recent attention is based on the Replica Exchange (RE)[11, 12] algorithm (also known as parallel tempering). To accomplish barrier crossings, RE methods simulate a series of replicas over a range of temperatures. Periodically, coordinates are exchanged using a Metropolis criterion[13] that ensures that at any given temperature a canonical distribution is realized. RE methods, particularly Replica Exchange Molecular Dynamics (REMD)[14], have become very popular for the study of protein biophysics, including peptide and protein folding[15, 16], aggregation[17, 18, 19], and protein-ligand interactions[20, 21]. Previous studies of protein folding appear to show a significant increase in the number of reversible folding events in REMD simulations versus conventional MD[22, 23]. Given the wide use of REMD, a better understanding of the RE algorithm and how it can be utilized most effectively for the study of protein folding and binding is of considerable interest.

The effectiveness of RE methods is determined by a complex of correlated factors, including the number of temperatures (replicas) that are simulated, their range and spacing, the rate at which exchanges are attempted and the kinetics of the system at each temperature. While the determination of “optimal” Metropolis acceptance rates and temperature spacings has been the subject of a variety of studies[12, 24, 25, 26, 27, 28, 29], the role played by the intrinsic temperature-dependent conformational kinetics which is central to

understanding RE has not received much attention. Recent work [30, 31, 29, 32] recognizes the importance of exploration of conformational space and the crossing of barriers between conformational states as the key limiting factor for the RE algorithm. Molecular kinetics can have a strong effect on RE beyond the entropic effects that have been discussed [30, 32], particularly if the kinetics does not have simple temperature dependence. It is known from experimental and computational studies that the folding rates of proteins and peptides can exhibit anti-Arrhenius behavior, where the folding rate *decreases* with increasing temperature[33, 34, 35, 36, 37, 38]. Different models have been proposed to explain the physical origin of this effect[39, 40]. To study the efficiency of RE under the context of anti-Arrhenius behavior is of considerable interest.

In chapter 2 and chapter 3, I will introduce two simplified models we built to simulate RE simulations of protein folding. These two models gave us great insight into the understanding of the mechanism of RE and will guide us to use RE in a more efficient way.

The other key challenge lies in the difficulty for an all-atom simulation to obtain meaningful information on the kinetics and pathways of the folding process. The typical timescale for a protein to fold is in magnitude of microseconds, which is much longer than the timescale of a conventional all-atom Molecular Dynamic(MD) simulation can reach in a reasonable computational time and have good statistics. 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 [41, 42, 43] If the process in question is activated, 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 an MC procedure is used to sample entire time-ordered paths connecting reactant and product wells in a well-defined manner[44]. 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 transition states.

A popular alternative takes advantage of heterogeneous distributed computing [45, 46] to enhance sampling by combining information from a large number of short molecular dynamics (MD) trajectories steered by rare events (Folding@Home). In a similar spirit, the ”milestoning” technique makes use of many short simulations that span two predefined critical points along a given reaction path [47]. While both approaches are powerful strategies, the former can introduce a bias towards fast events in the ensemble average of the reactive trajectories [48], while the latter is limited to a single reaction path that must be specified in advance. Thus, neither of these approaches can be used to effectively study systems that may have multiple pathways and transition states.

A related set of methods for obtaining kinetic information are based on the use of stochastic dynamics on a free energy landscape [49, 50, 51, 52, 53, 54]. They are based 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 to study the kinetics of the system by low-dimensionality Langevin simulations. While various strategies have been proposed to discover good reaction coordinates in complex systems [55, 56, 57] , 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 multiple transition states.

An additional 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 [58], and the kinetics on this graph is often assumed to have Markovian behavior [59, 60, 61, 62, 63]. This approach is particularly well suited for reduced lattice models, and was first introduced in that context [59]. For systems with a continuous state space, some form of discretization is required. This can be done by clustering based on chosen reduced coordinate [58, 61], though the clusters must be chosen carefully so as to satisfy the Markovian condition [62, 63, 64, 65]. Alternatively, the

discretization can be based on an analysis of the minima and/or saddle points of the energy surface [60, 66, 67], 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 [68]. The location of all minima or saddle points, however, can be a serious challenge for high-dimensional systems, though it has shown that this is possible for peptide systems [67, 69]. A hybrid approach has also been proposed that makes use of molecular dynamics to infer local transition regions to build disconnectivity graphs [70].

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 [51, 56], 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.

Powerful generalized ensemble methods [71] such as replica exchange molecular dynamics (REMD) [72] have been developed which enhance the ability to obtain accurate canonical populations in complex systems by increasing sampling efficiency. However, since REMD involves temperature swaps between MD trajectories, it is not straightforward to obtain kinetic information from such simulations. [63, 73, 54]. Our laboratory has made use of a kinetic network model [74] in which the nodes correspond to molecular conformations from REMD simulation trajectories, and the edges are derived from an ansatz based on structural similarity. While this model was shown to yield physically plausible kinetics [74] , 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.

We are going to 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 multi-dimensional configurational space of the system by running 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) [75], allowing us to obtain correct thermodynamic averages from the RE samples over all simulation temperatures. We then use short-time local dynamics to derive drift velocities and diffusion coefficients on a suitably chosen reduced coordinate. The network topology and microscopic rate parameters can then be adjusted recursively to match the drift velocities and diffusion coefficients derived from the network simulations to those derived from 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 slower kinetic processes than would be accessible using conventional MD. In chapter 4, I will demonstrate our approach using the folding like two-dimensional potential constructed in chapter3 and discuss generalizations to the more complex energy landscapes of atomic-level protein simulations.

Chapter 2

Simulating Replica Exchange Simulations of protein folding with a kinetic network model

To understand to what extent the efficiency of replica exchange will be affected by the kinetics of the biomolecular system and the replica exchange parameter set, simplified model is a good choice based on two reasons: first, it will take a much shorter time for the system to converge so that we can run it for a large ensemble of different instances to obtain good statistics for the data; second, we have much more freedom in controlling the parameters of the system and of the replica exchange setup and we can observe the system's behavior under extreme conditions which could separate the effect of different parameters.

In this chapter, we investigate the impact of simulation parameters and anti-Arrhenius kinetics on the RE method. Because RE simulations of protein systems that display anti-Arrhenius behavior are difficult to converge, we developed a kinetic network RE (NRE) model that allows us to simulate the RE algorithm of two-state protein folding. This network model reduces the atomic complexity of the system to a set of discrete conformational states that evolve in continuous time according to Markovian kinetics for both conformational transitions and exchange between replicas.

Kinetic network model has been used to improve computational efficiency by discretizing the state space and constructing rules for moving among those states. The resulting scheme can be represented as a graph or network[76]. The kinetics on this graph is assumed to be stochastic, leading to a Markovian model for the time dependence of the populations of the various states[77, 78, 79, 80, 81, 82]. Similar schemes have been constructed based on the output of more conventional MD simulations (often after clustering and choosing a

reaction coordinate)[76, 79, 80, 81, 83] or based on an analysis of the minima and/or saddle points of the energy surface[82, 84, 85].

The NRE model studied here does not capture many of the complexities of the ”real” molecular simulation. For example, it does not have finite-width energy distributions, and the kinetics of atomic-level simulations are likely to exhibit various kinds of non-Markovian behavior. However, it does capture many of the essential features of RE and allows us to study these fundamental aspects of the algorithm at low computational cost and in a controlled setting. This allows us to separate the interacting parameters and study their effects on the simulation individually. Given that NRE is an idealized version of RE, many of the limitations in the convergence rates and efficiency observed with NRE will likely also be present in full atomic-level RE simulations (in addition to further limitations created by the complexities of the atomic-level simulations), allowing us to identify promising avenues of inquiry for future atomic-level simulations.

2.1 Introduction to Replica Exchange Method

Let us consider an original system of N atoms with Hamiltonian $H(X)$, where X is a state of the system, i.e. a point in the phase space. In the canonical ensemble at temperature T, the equilibrium probability of state X follows Boltzmann distribution:

$$P_{eq}(X,T) = \frac{\exp[-\beta H(X,T)]}{Z(T)},$$

where $\beta = (k_B T)^{-1}$ is the inverse temperature, and $Z(T)$ is the partition function of the system. If we simulate this system using the conventional Monte Carlo method (MC), in order for the simulation to converge to the equilibrium distribution, it is sufficient to impose the detailed balance condition on the transition probability $w(X \to X')$:

$$P_{eq}(X,T)w(X \to X') = P_{eq}(X',T)w(X' \to X).$$

By the Metropolis criterion,

$$w(X \to X') = \begin{cases} 1 & \Delta_0 \leq 0 \\ \exp(-\Delta_0) & \Delta_0 > 0 \end{cases},$$

where $\Delta = \beta[H(X', T) - H(X, T)]$. If X is a local-minimum-energy state, at the neighborhood of X, $H(X') > H(X)$. At low temperature, the transition probability $w(X \to X') \ll 1$, but the probability of pulling the new sampling state back to the local-minimum state X is $w(X' \to X) \approx 1$. In this case, the sampling state has a negligibly small probability to leave the neighborhood of state X. This is an example of the simulation being trapped in a local energy minimum. Suppose we have a combined system which consists of M non-interacting replicas, each replica is the original system contacting with different heat bath of temperature T_m ($T_1 < T_2 < \ldots < T_M$). A state of this extended ensemble is specified by a joint configuration of M replicas $\{X^{(1)}_{m(1)}, X^{(2)}_{m(2)}, \ldots, X^{(M)}_{m(M)}\}$, where $X^{(i)}_{m(i)}$ stands for the configuration of replica i at temperature $T_{m(i)}$ and $m(i)$ is a permutation of replica label $i = 1, 2, \ldots, M$.

If M replicas are distinguishable and non-interacting, the equilibrium distribution of the extended state $\{X^{(1)}_{m(1)}, X^{(2)}_{m(2)}, \ldots, X^{(M)}_{m(M)}\}$ is

$$\begin{aligned} P_{eq}(\{X^{(1)}_{m(1)}, X^{(2)}_{m(2)}, \ldots, X^{(M)}_{m(M)}\}) &= \frac{1}{M!} \prod_{i=1}^{M} P_{eq}(X^{(i)}, T_{m(i)}) \\ &= \frac{1}{M!} \prod_{i=1}^{M} \frac{\exp(-\beta_{m(i)} H(X^{(i)}, T_{m(i)}))}{Z(T_{m(i)})}, \end{aligned}$$

where $M!$ is the normalizing constant (because of the permutation of M configurations of each replica) , and

$$P_{eq}(X, T) = \frac{\exp(-\beta H(X, T))}{Z(T)}$$

is the canonical equilibrium distribution of the original system (single replica).

To simulate the extended ensemble, in addition to the local MC move within each replica, we introduce a temperature exchange between two replicas. e.g., we exchange

the temperatures of replica i and j: $\{X\} = \{\ldots, X^{(i)}_{m(i)}, \ldots, X^{(j)}_{m(j)}, \ldots\} \rightarrow \{X'\} = \{\ldots, X^{(j)}_{m(j)}, \ldots, X^{(i)}_{m(i)}, \ldots\}$. The detailed balance condition will be $P_{eq}(\{X\})w(\{X\} \rightarrow \{X'\}) = P_{eq}(\{X'\})w(\{X'\} \rightarrow \{X\})$. The equation can be simplified when the Hamiltonian of each replica is independent of temperature, i.e. $H(X,T) = H(X)$. By using Metropolis criterion, we have

$$w(\{X\} \rightarrow \{X'\}) = \begin{cases} 1 & \Delta \leq 0 \\ \exp(-\Delta) & \Delta > 0 \end{cases},$$

where $\Delta = (\beta_{m(j)} - \beta_{m(i)})(H(X^{(i)}) - H(X^{(j)}))$.

As shown in Fig. 2.1, rough energy landscape of a protein trap simulations in a local minimum when conventional sampling methods are used, like Monte Carlo methods or Molecular Dynamics methods. When using Replica Exchange method, low temperature replica borrows fast kinetics from high temperature replica and speeds up the equilibration of the system at low temperature. This is in principle the fundamental mechanism of how and why replica exchange works in enhancing the sampling efficiency. When applied in real molecular systems, however, the efficiency of RE will be affected by at least two factors: the values of RE parameters(e.g. Frequency of attempting replica exchange, number of replicas, the temperature range and distribution for the replicas, etc.) and the kinetic property of the system. Especially for the latter, if a system does not have fast kinetics at high temperature, it is not going to do any good using replica exchange in the first place. In chapter 2 and chapter 3, we constructed a discrete network model and a continuous two-dimensional potential to simulate Replica Exchange simulation and show in great details why and how these two factors affect the efficiency of RE.

2.2 Construction of the network model

In order to isolate some of the essential features of the RE algorithm, we construct a kinetic network model of RE (NRE) which we can use to study the effects of the parameters of the

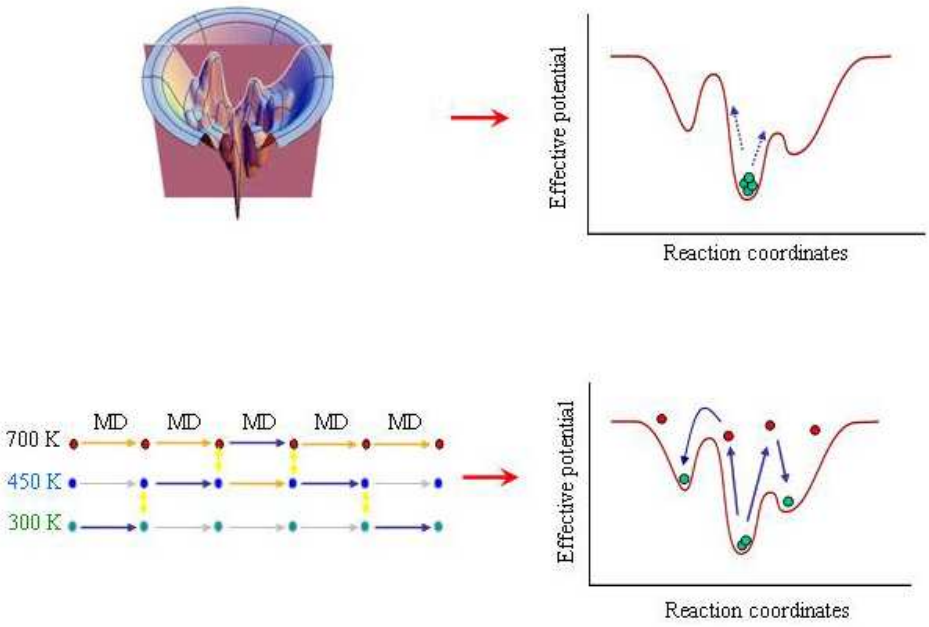

Figure 2.1: Rough energy landscapes of protein folding trap simulations in a local minimum when conventional sampling methods are used(e.g. Monte Carlo or Molecular Dynamics). Using Replica Exchange method, the low temperature replica borrows fast kinetics from high temperature replica to help escaping the local minimum at low temperature.

model on efficiency and convergence. We consider a system in which the configurational space can be partitioned into two macrostates of interest separated by a free energy barrier that makes transitions between the conformations an activated process. Motivated by protein folding, we call these macrostates F and U (for "Folded" and "Unfolded"). Transitions between F and U in a (non-RE) MD or kinetic MC simulation can be approximated by a Poisson process in which the waiting times between folding and unfolding transition events are exponentially distributed random variables with mean equal to the reciprocal of the folding or unfolding rates, respectively.

If the transition events are Markovian, then we can represent the simultaneous behavior of two non-interacting replicas in terms of the four composite states $\{F_1F_2, F_1U_2, U_1F_2, U_1U_2\}$. In each symbol, the first letter is the configuration of replica 1, the second letter is the configuration of replica 2, and the subscripts are the temperature of each replica. Therefore F_1U_2 represents the composite state that replica 1 at temperature T_1 is folded, while replica 2 at temperature T_2 is unfolded. The kinetics in the composite state space can be represented as a continuous-time Markov process with discrete states[86].

The four-state composite system corresponding to non-interacting replicas can be extended to create a discrete-state model of replica exchange by introducing temperature exchanges between replicas. For example, suppose the current state is F_1U_2. After a successful temperature exchange, replica 1 is at T_2 and replica 2 is at T_1 and the new state can be represented as F_2U_1. The introduction of temperature exchange therefore creates four additional states, leading to the 8-state system $\{F_1F_2, F_1U_2, U_1F_2, U_1U_2, F_2F_1, F_2U_1, U_2F_1, U_2U_1\}$. These states are arranged into two sub-networks defined by the "horizontal" folding and unfolding transitions, which are connected to each other by "vertical" temperature exchange transitions, forming a cubic network (Figure 2.2). In general, the network for an N-replica system consists of $N!$ sub-networks, each of which has 2^N states connected by folding/unfolding transitions. The model description in this section will focus primarily on the 2-replica case; all of the details can be easily generalized to the case of N replicas.

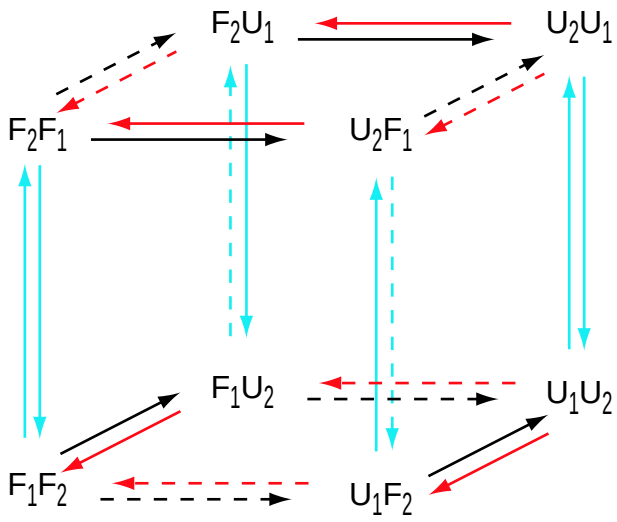

Figure 2.2: The kinetic network of the composite states corresponding to the simplified replica exchange model with two replicas. The state labels represent the conformation (letter) and temperature (subscript) for each replica. For example, F_2U_1 represents the state in which replica 1 is folded and at temperature T_2, while replica 2 is unfolded and at temperature T_1. Red and black arrows correspond to folding and unfolding transitions, respectively, while the temperature at which the transition occurs is indicated by the solid and dashed lines (for T_2 and T_1, respectively). The cyan arrows correspond to temperature exchange transitions, with the solid and dashed lines denoting transitions with rate parameters α and $w\alpha$, respectively.

We require that the equilibrium populations of the states be such that the canonical ensemble is recovered at each temperature. This is the case if the equilibrium populations are proportional to the product of the equilibrium populations for the two-state systems, e.g.

$$P_{eq}(F_1U_2) = \frac{1}{2}P_{eq}(F_1)P_{eq}(U_2) = \frac{1}{2}\frac{k_{f1}k_{u2}}{(k_{f1}+k_{u1})(k_{f2}+k_{u2})},$$

where the factor of $1/2$ accounts for the presence of the two equivalent manifolds. For these probabilities to be preserved under temperature exchanges, it is sufficient that detailed balance is satisfied, e.g. the transition probabilities $w(F_1U_2 \rightarrow F_2U_1)$ and $w(F_2U_1 \rightarrow F_1U_2)$ satisfy $P_{eq}(F_1U_2)w(F_1U_2 \rightarrow F_2U_1) = P_{eq}(F_2U_1)w(F_2U_1 \rightarrow F_1U_2)$, or

$$\frac{w(F_1U_2 \rightarrow F_2U_1)}{w(F_2U_1 \rightarrow F_1U_2)} = \frac{k_{f2}k_{u1}}{k_{f1}k_{u2}} \equiv w. \tag{2.1}$$

If the equilibrium favors the folded state at T_1 and the unfolded state at T_2, then $w < 1$. The ratios of forward and reverse transition probabilities for $F_1F_2 \rightleftharpoons F_2F_1$ and $U_1U_2 \rightleftharpoons U_2U_1$ are equal to one, as interchange of temperatures does not change the equilibrium populations.

In atomic-level RE simulations, temperature exchange attempts are usually made periodically in time, i.e. the MC or MD evolution is interrupted, temperature swap proposal(s) are made, and the proposals are either accepted or rejected[14, 16]. In keeping with the continuous-time nature of our network model, we simulate the effect of temperature exchanges by introducing an additional rate parameter α which controls the overall scaling of the temperature exchange rate relative to the folding and unfolding rates. We set the forward and reverse rates of the $F_1F_2 \rightleftharpoons F_2F_1$ and $U_1U_2 \rightleftharpoons U_2U_1$ "reactions" equal to α, while the other rates are set to α or $w\alpha$ (Figure 2.2) as required by detailed balance (Eq. 2.1), and where we choose $w < 1$. For example, the states U_1F_2 and U_2F_1 differ in population, with U_2F_1 being more populated if the equilibrium favors the folded state at T_1 and the unfolded state at T_2. We therefore set the $U_1F_2 \rightarrow U_2F_1$ "reaction rate" equal to α, and the reverse rate equal to $w\alpha$, where w is defined in Eq. 2.1.

The NRE model can be simulated using a standard method for continuous time Markov processes with discrete states[86], also known as the "Gillespie algorithm". The algorithm remains efficient even when the number of replicas is large (e.g. 20 replicas, corresponding to 10^{24} states) due to the fact that each state is connected to a small number of neighboring states (those connected by single temperature exchanges involving neighboring temperatures and folding/unfolding transitions of each replica).

The convergence or efficiency of a simulation is monitored by measuring $N_{\mathrm{TE}}(\tau|T_1)$, the number of "round-trip" transitions between the U and F states, conditional on the temperature of interest T_1 that occur in a given observation time τ. In the context of the network model, suppose that we follow replica 1, and at a given time the system is in a state where that replica is folded at temperature T_1 (e.g. F_1F_2). We then wait for the first occurrence of a state in which replica 1 is unfolded at T_1 (e.g. U_1F_2), and then for the first occurrence of a state in which that replica is folded again at T_1 (e.g. F_1F_2). At this point, we say that a transition event has occurred. Conceptually, a transition event is a transit of a given replica from one conformation at low temperature to the other conformation at low temperature and back again regardless of route, i.e. whether it was the result of a direct barrier crossing at T_1 or indirectly via a barrier crossing at T_2 combined with temperature exchanges. The number of transitions as defined corresponds to the number of "reversible folding" events studied in all-atom simulations of peptide systems[22, 23].

2.2.1 Thermodynamic model for anti-Arrhenius behavior

The Arrhenius equation relates a reaction rate k to the temperature:

$$k(T) = Ae^{-\Delta G^{\dagger}(T)/k_BT} = Ae^{-(\Delta E^{\dagger}(T) - T\Delta S^{\dagger}(T))/k_BT}, \tag{2.2}$$

where $\Delta G^{\dagger}(T)$ is the free energy of activation. The temperature dependence of the reaction rate is customarily described by means of the Arrhenius plot, the plot of $\ln k(T)$ with respect to $1/T$. The slope of $\ln k(T)$ in the Arrhenius plot is proportional to the activation

energy, $\Delta E^{\dagger}(T)$, at temperature T. When the activation energy is temperature independent the Arrhenius plot appears as a line of constant slope. Moreover, if the activation energy is positive, the reaction rate increases with increasing temperature. This behavior is referred to as normal Arrhenius behavior. When the activation energy is negative, however, increasing the temperature causes the rate to decrease. This non-intuitive phenomenon sometimes observed in protein folding kinetics[33, 34, 35, 36, 37, 38] is referred to as anti-Arrhenius behavior. In these circumstances the transition state is energetically favored but entropically disfavored with respect to the reactants.

Often protein folding rates follow normal Arrhenius behavior at low temperatures, switching to anti-Arrhenius behavior at higher temperatures. This mixed behavior can be understood in terms of a constant activation heat capacity model in which the activation energy and entropy vary linearly with respect to the temperature and its logarithm, respectively[87, 34] :

$$\Delta E^{\dagger}(T) = \Delta E^{\dagger}(T_0) + (T - T_0)\Delta C_p^{\dagger} \tag{2.3}$$

$$\Delta S^{\dagger}(T) = \Delta S^{\dagger}(T_0) + \ln(T/T_0)\Delta C_p^{\dagger} \tag{2.4}$$

where $\Delta C_p^{\dagger} < 0$ is the activation heat capacity which is assumed here to be independent of temperature. Summing Eqs. 2.3 and 2.4, we obtain the expression for $\Delta G^{\dagger}(T)$ corresponding to this model. The Arrhenius plots for the unfolding and folding rates, $k_u(T)$ and $k_f(T)$ used in this work, that result from inserting this expression in Eq. 2.2, setting $\ln A/s^{-1} = 22$, $T_0 = 300$K, and $\Delta E^{\dagger}(T_0)$, $\Delta S^{\dagger}(T_0)$, and $\Delta C_p^{\dagger}$ to be 2 kcal/mol, -0.01 kcal/mol/K, and -0.025 kcal/mol/K for folding, and 8.5 kcal/mol, 0.008 kcal/mol/K, and 0 kcal/mol/K for unfolding, respectively, are shown in Figure 2.3. For the case of Arrhenius folding (Figure 2.3 dashed line), the parameters are identical with the exception that $\Delta C_p^{\dagger}$ for folding is zero. The unfolding rate follows normal linear Arrhenius behavior, whereas the anti-Arrhenius folding rate decreases with increasing temperature above $T^* = 380$ K (the temperature at which the activation energy for folding is zero and the

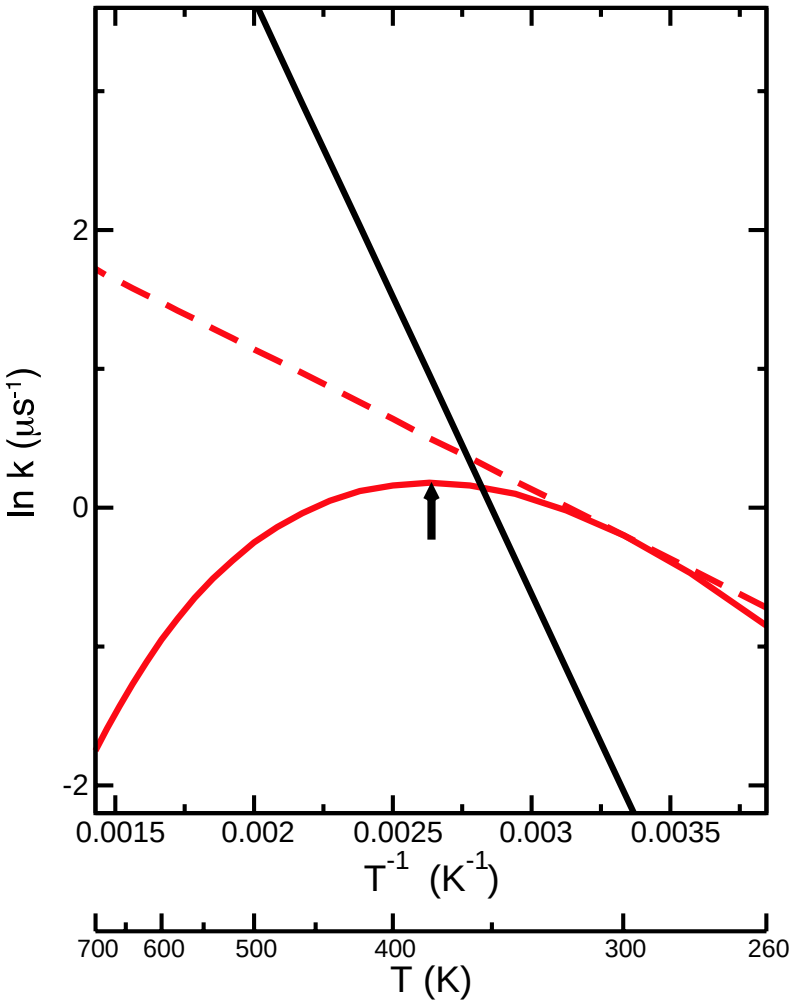

Figure 2.3: Arrhenius plot of the folding and unfolding rates from a thermodynamic model for the temperature dependence of protein folding rate constants. Black line corresponds to unfolding rate, while red lines correspond to the folding rates. The solid line is for the $\Delta C_p^\dagger \neq 0$ case displaying anti-Arrhenius behavior, while the dashed line corresponds to the same parameters with $\Delta C_p^\dagger = 0$. The arrow indicates the temperature T^* at which the folding rate is maximal (≈ 380 K).

folding rate is maximal). The general behavior of $k_u(T)$ and $k_f(T)$ shown in Figure 2.3 is typical for experimentally determined peptide folding kinetic rates[33, 35, 38].

2.3 Results and Discussion

2.3.1 Convergence efficiency of non-RE simulations

When considering questions of efficiency, it is often useful to compare results to that of a well-understood reference. In the case of RE simulations, we choose a single-temperature uncoupled MD or kinetic MC simulation as the reference. If we assume (as in NRE) that kinetics over a discretized state space is Poisson, then the convergence behavior of the single-temperature simulation can be determined analytically. Let us consider a system with rates

$$F \underset{k_f}{\overset{k_u}{\rightleftharpoons}} U$$

and suppose that we are interested in estimating the equilibrium population in the F state (the "fraction folded"). In molecular simulations, this is typically estimated by allowing the system to evolve for a certain amount of time τ and calculating the fraction of time spent in the F state:

$$S(\tau) = \frac{1}{\tau} \int_0^\tau \delta_F(t) dt, \tag{2.5}$$

where $\delta_F(t)$ is an indicator function that is 0 if the system is in state U at time t and 1 if it is in state F. If the system is Poisson, then $S(\tau)$ is the random variable corresponding to the normalized time integral of the "telegraph process", which is the Markov process in which the system alternates between states 0 and 1 with exponential residence times[86]. The moments of the time integral of the telegraph process can be determined analytically[86], and can be normalized by $1/\tau$ to obtain the moments of $S(\tau)$. In particular, the mean of $S(\tau)$ is given by

$$\langle S(\tau) \rangle_F = \frac{k_f}{k_u + k_f} + \frac{k_u}{\tau (k_u + k_f)^2} \left(1 - e^{-(k_u + k_f)\tau}\right) \tag{2.6}$$

or

$$\langle S(\tau)\rangle_U = \frac{k_f}{k_u + k_f} - \frac{k_f}{\tau(k_u + k_f)^2}\left(1 - e^{-(k_u+k_f)\tau}\right) \tag{2.7}$$

depending on whether the system began in state F or U at time $t = 0$, respectively. These are of interest because they tells us how quickly the system equilibrates.

Equations 2.6 and 2.7 show that the mean of $S(\tau)$ approaches the true fractional population as $\tau \to \infty$, and the second term in each equation represents the mean deviation from the correct value. The magnitude of this bias depends strongly on the starting state: e.g. beginning in F leads to much smaller bias if the system's equilibrium strongly favors F. In a molecular simulations, one normally does not know *a priori* where the equilibrium lies, and therefore which is the more favorable starting state. One can account for this uncertainty using the average absolute bias

$$\frac{1}{2}\left(\langle S(\tau)\rangle_F - \langle S(\tau)\rangle_U - \frac{2k_f}{k_u + k_f}\right) = \frac{1}{\tau(k_u + k_f)}\left(1 - e^{-(k_u+k_f)\tau}\right) \tag{2.8}$$

corresponding to the average over choosing the starting state to be U or F with equal probability. The average absolute bias depends inversely on the rates only via their sum, and becomes negligible if τ is large relative to $(k_u+k_f)^{-1}$. Therefore, the bias is dominated by the fastest rate, in the sense that if k_u and k_f are of different magnitudes, changes in the smaller of the two will have very little effect on the convergence compared to changes in the larger rate. The origin of this can be most easily seen in the limit where $k_f \gg k_u$. If we begin in F, then even if no transition events occur we will have little bias, since the true value of the fraction folded is very close to 1. Alternatively, if we begin in U, then we will be very likely to quickly see a folding event (provided that τ is not too small), again leading to small bias. The key observation is that, for a non-RE simulation, the convergence is dominated by the fastest rate, and in some circumstances it is not necessary to have many "round-trip" transitions between the states in order to obtain converged results.

2.3.2 Convergence efficiency of the kinetic network model for large α limit

We first examine the behavior of NRE for the simplest possible case: two replicas where the rate of temperature exchanges is large compared to the folding/unfolding kinetics. The condition that α be very large relative to the "molecular" kinetic rates simplifies the problem, since in that limit the behavior will be independent of the precise choice of α and will depend only on the (temperature-dependent) folding and unfolding rates. Since the energy distributions in NRE are temperature independent δ-functions, there is no intrinsic penalty for having the temperature difference between the replicas be very large. Therefore, we fix T_1 at 300K, and sweep T_2 over the range 300K to 700K. We wish to see if there is a specific T_2 which gives optimal convergence. In Figure 2.4 we show the estimates of the "fraction folded" $S_1(\tau)$ averaged over many independent simulations. The fraction folded $S_1(\tau)$ is defined as the fraction of time spent in the F state at low temperature:

$$S_1(\tau) = \frac{1}{\tau}\int_0^\tau \delta_{F1}(t)dt, \tag{2.9}$$

where $\delta_{F1}(t)$ is an indicator function that is 1 if the system is in one of the four composite states in which the replica at T_1 is folded (F_1F_2, F_1U_2, F_2F_1, or U_2F_1), and 0 otherwise. Since the time τ used for these simulations is not large relative to the equilibration time of the system, there is a significant deviation of $\langle S_1(\tau)\rangle$ from the correct value (indicated by the horizontal dotted line), and the distance from the curves to the dotted line represents the bias. In the Arrhenius ($\Delta C_p^\dagger = 0$) case, the bias decreases monotonically with T_2. For the protein folding model having anti-Arrhenius behavior ($\Delta C_p^\dagger \neq 0$), there is a clear minimum in the bias at $T_2 \approx 440$ K. Thus, unlike the purely Arrhenius case, there is an unambiguous optimal high temperature.

In order to investigate the origin of this optimal temperature, We have measured the number of conformational transitions for the NRE model using both Arrhenius and anti-Arrhenius models for the folding and unfolding rates with various choices of the number

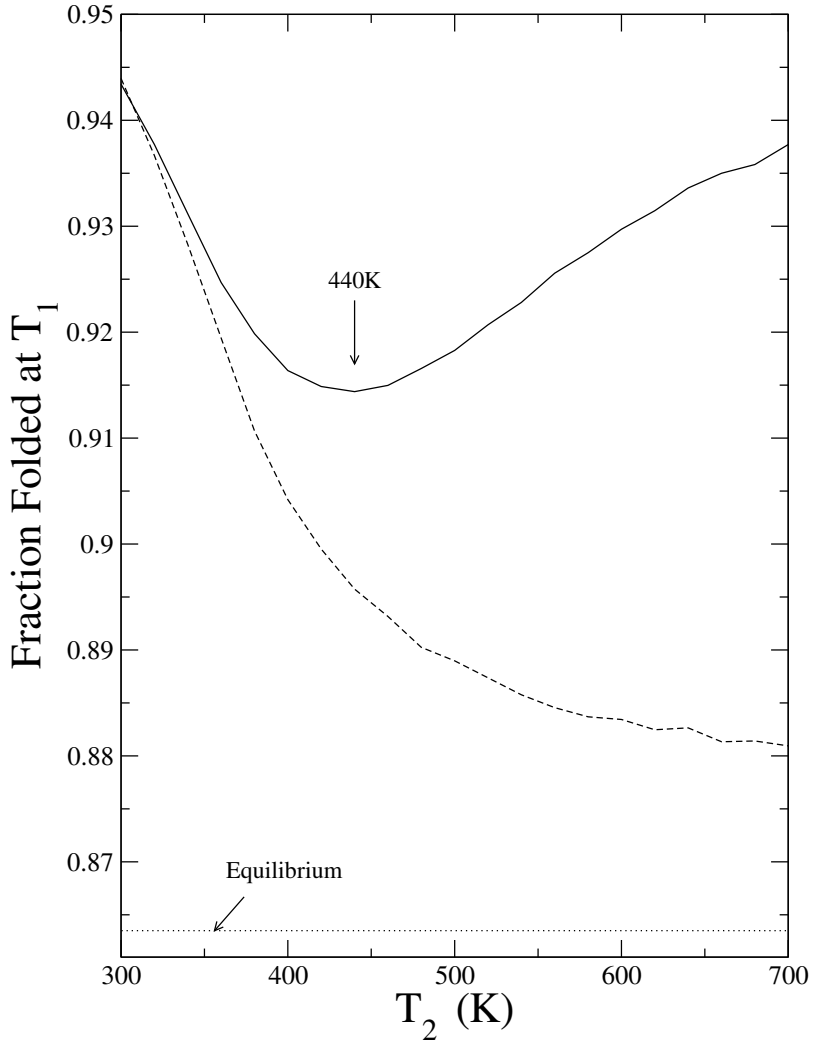

Figure 2.4: Estimates of the relative population of the F conformation at temperature $T_1 = 300$ K for a finite simulation time. The temperature of replica 1 was held fixed at 300 K, while T_2 (of replica 2) is swept from 300 K to 700 K. The temperature exchange parameter α was set to 2 ns^{-1}. For each individual T_2, the system was simulated for $\tau = 1.25$ μs beginning in the state F_1F_2 at time $t = 0$ and the fraction folded at T_1 $S_1(\tau)$ was calculated. This was repeated 50,000 times, and the resulting $S_1(\tau)$ values were averaged and the results are plotted. The solid line corresponds to the anti-Arrhenius folding rates ($\Delta C_p^\dagger \neq 0$), while the dashed line corresponds to the Arrhenius rates ($\Delta C_p^\dagger = 0$) (Figure 2.3). The true fraction folded at $T_1 = 300$ K is the same for both the Arrhenius and anti-Arrhenius models and is indicated by the dotted line. The temperature at which the bias is minimized for the anti-Arrhenius model (≈ 440 K) is indicated by the arrow.

of replicas, their temperatures, and the temperature exchange rate parameter α. The goal of these calculations is to study factors that affect the increased efficiency that RE can provide.

We define the efficiency in the context of NRE to be the total number of transition events divided by the number of replicas $N_{\mathrm{TE}}(\tau|T_1)/N$. We make several general observations. First of all, increasing the total temperature range for a given number of replicas can degrade the efficiency of reversible folding if the kinetics is anti-Arrhenius (Figure 2.5A). On the contrary, for the purely Arrhenius case, both the folding and unfolding rates increases as temperature goes up, results in no optimum temperature for the efficiency(dashed line).

In order to understand this behavior, we first examine the behavior of NRE for the simple case of two replicas ($N = 2$), where the rate of temperature exchanges is large compared to the folding/unfolding kinetics. The condition that α be very large relative to the conformational kinetic rates simplifies the problem, since in that limit the behavior is independent of the precise choice of α and depends on the (temperature-dependent) folding and unfolding rates. We fix T_1 at 300K, and sweep T_2 over the range 300K to 700K. In Figure 2.6A we show the dependence of $N_{\mathrm{TE}}(\tau|T_1)/N$ normalized by the number of replicas as a function of T_2 for the anti-Arrhenius kinetic model ($\Delta C_p^\dagger \neq 0$).$N_{\mathrm{TE}}(\tau|T_1)/N$ indicates the convergence efficiency of the system. We see that, for the two-replica system, $N_{\mathrm{TE}}(\tau|T_1)/N$ is small at low and high T_2, and reaches a maximum near 440 K (dashed black line).

The number of transition events for an uncoupled, non-RE simulation is easy to predict. If the kinetics is Poisson, then the mean lifetime in each basin is k^{-1}, where k is the rate for leaving the basin. In order to make a round-trip starting from F, for example, we must wait on average k_u^{-1} before jumping to U, and then another k_f^{-1} before jumping back to F. Therefore the rate of transition events is given by the harmonic mean of the folding and unfolding rates $(k_u^{-1} + k_f^{-1})^{-1}$, and the number of transitions is that rate multiplied by the total observation time. The harmonic mean is dominated by the smallest rate, agreeing

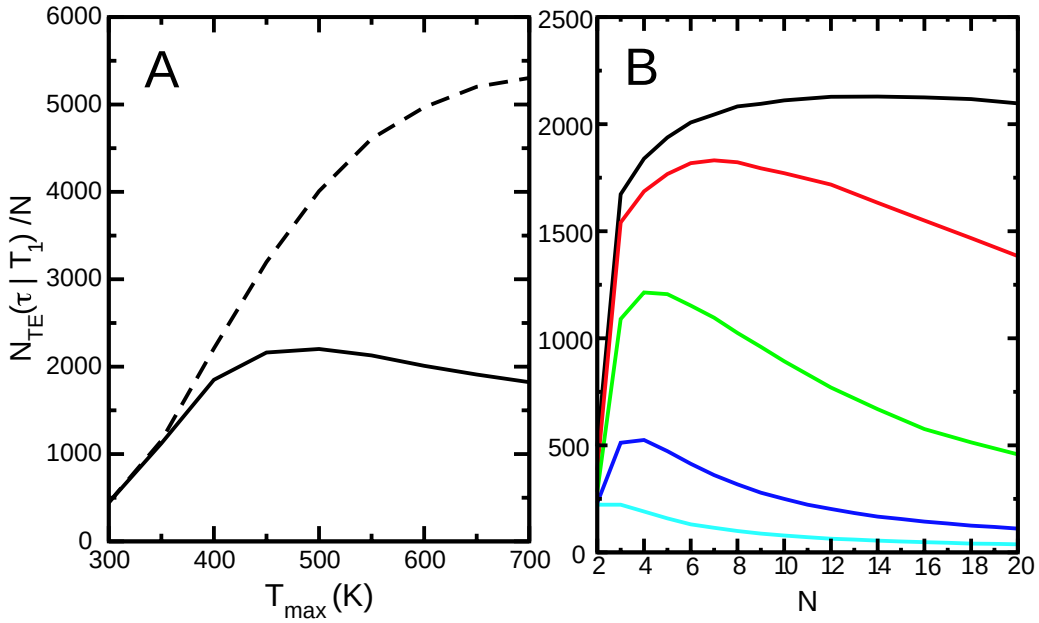

Figure 2.5: Number of transition events in NRE simulations (normalized by the number of replicas) for various temperature ranges, exchange rates α, and number of replicas N. In all cases, the system was simulated for $\tau = 4\ \mu s$. For the simulations in (A), α was set to 1000 μs^{-1}, the dashed and solid lines correspond to Arrhenius and anti-Arrhenius kinetics, respectively, and six replicas were exponentially distributed between 300 K and T_{max}. The simulations in (B) were performed with anti-Arrhenius rates, N replicas exponentially distributed from 300 K to 700 K, and α values of 10000 μs^{-1} (black), 1000 μs^{-1} (red), 100 μs^{-1}(green), 10 μs^{-1}(blue) and 1 μs^{-1}(cyan).

with our intuition that the number of transitions is determined by the rate limiting step.

In the case of coupled trajectories such as in the NRE model, the dependence of the number of transitions on the rates is not as obvious, however it can be easily estimated by simulation. In Figure 2.6A we show the dependence of $N_{\mathrm{TE}}(\tau|T_1)/N$ as a function of T_2 for the anti-Arrhenius kinetic model. We see that $N_{\mathrm{TE}}(\tau|T_1)/N$ is small at low and high T_2, and reaches a maximum near 440 K (Figure solid black line). In fact, the $N_{\mathrm{TE}}(\tau|T_1)/N$ obtained by simulation in the large α limit is very well approximated by the "arithmetic mean of harmonic means"

$$N_{\mathrm{TE}}(\tau|T_1)/N \approx \frac{\tau}{N}\left[(k_{f1}^{-1}+k_{u1}^{-1})^{-1}+(k_{f2}^{-1}+k_{u2}^{-1})^{-1}\right]$$

(Figure 2.6 black dashed line). These results suggest that the convergence of NRE is limited by the rate at which round-trips between basins occur, and that the convergence rate is therefore strongly dependent on the slowest rates. This is very different from the uncoupled, non-RE case discussed above, which is dominated by the fastest rate. The system must sample all basins more than once in order to accurately estimate populations, and the convergence at T_1 will be limited by the number of transitions $N_{\mathrm{TE}}(\tau|T_1)/N$.

Next, we examine how the number of replicas affects the convergence as monitored by the number of transition events. In Figure 2.6A we examine whether a third replica results in an improvement over the optimum behavior with two replicas. To do this, we fix T_1 at 300 K, T_3 at 440 K (the two-replica optimum), and scan T_2 from 300 K to 700 K (i.e. we do not require $T_1 < T_2 < T_3$). We see in Figure 2.6A (solid green line) that the number of transitions per replica again reaches a maximum near $T_2 \approx 440$ K, corresponding to the case where one replica is at the temperature of interest (300 K), while the other two are both placed at the "optimal" temperature of 440 K. As in the two-replica case, $N_{\mathrm{TE}}(\tau|T_1)/N$ is very well-approximated by the average of the harmonic means of the rates at all three temperatures (Figure 2.6A dashed green line). The relevant question is whether the addition of the third replica is an improvement over two. It is important in this regard to distinguish the convergence rate from the computational efficiency of the simulation. In the cases seen

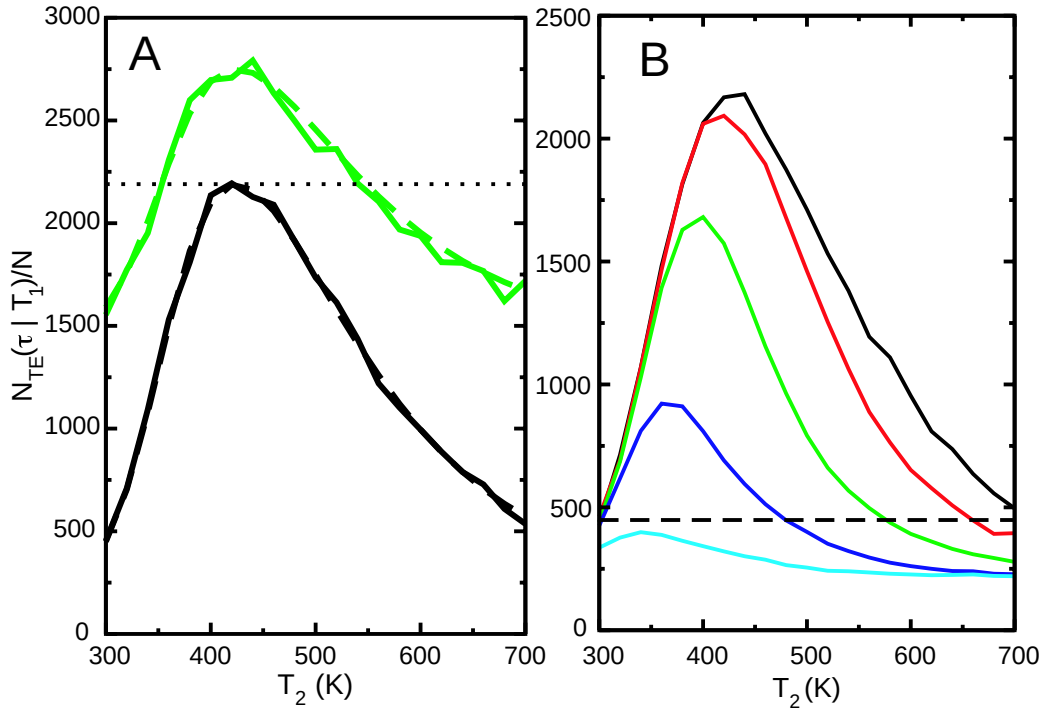

Figure 2.6: Number of transition events per replica in NRE simulations using the anti-Arrhenius folding rates for a simulation time $\tau = 4$ ms conditional on temperature $T_1 = 300$ K, while T_2 is scanned from 300 K to 700 K. (A) Black and green solid lines: simulation results for two-replica and three replica systems (with $T_3 = 440$ K), respectively. Black and green dashed lines: number of transition events predicted using the average of harmonic means for two and three replicas, respectively. All simulations were performed with $\alpha = 10$ ns^{-1}. (B) Results for two-replica NRE simulations using the anti-Arrhenius folding rates and α values of 10 ns^{-1} (black solid), 1 ns^{-1} (red), 100 μs^{-1} (green), 10 μs^{-1} (blue), and 1 μs^{-1} (cyan). The black dashed line corresponds to the predicted number of transitions for a single, uncoupled simulation at T_1.

in Figure 2.6A, the total number of transition events (*not* normalized by the number of replicas) is larger for three replicas than the maximum total number of transition events for two replicas, and therefore we expect the convergence to be better. In general, adding an additional replica will always improve overall convergence, since the additional transition pathways opened up will always have a positive contribution to the total number of transition events. However, the computational efficiency of NRE as measured by $N_{\mathrm{TE}}(\tau|T_1)/N$ of the three-replica simulation is improved relative to the two-replica simulation only if the additional temperature T_2 has values between 350 and 550 K (Figure 2.6A black dotted line). While the addition of a replica always improves convergence, it improves efficiency only if the harmonic mean of the rates at the additional temperature is large relative to the harmonic means of the other replicas. If not, then the presence of the additional slow paths will reduce the efficiency. For the general case of NRE with N replicas, we expect that, in the large α limit, optimal efficiency (and convergence) will be obtained when one replica is at the temperature of interest, and all of the other replicas are placed at the temperature which maximizes the harmonic mean of the folding and unfolding rates. Thus, the replica with the largest harmonic mean sets a "speed limit" for the amount of efficiency improvement that an RE simulation can have over an uncoupled simulation run for the same amount of CPU time. The addition of replica $N + 1$ will increase the efficiency only if the harmonic mean at the new temperature is greater than the average of the harmonic means of the original N replicas.

2.3.3 Convergence efficiency of the kinetic network for finite α

In the results described above, the rate of temperature exchanges is so large that convergence is limited only by the rates of conformational transitions at each temperature. When α is comparable to or smaller than the rates of conformational transitions, the waiting time for a temperature exchange to occur becomes comparable to or even larger than the time scale of configuration changes within each replica. Therefore, there can be multiple folding

or unfolding events at higher temperatures before any of these events are transmitted to the temperature of interest. These events are "lost" and make no contribution to the number of transition events at low temperature. Therefore, in the NRE model (where conformational transitions are instantaneous and strictly Markovian), the optimal convergence (and efficiency) is achieved in the limit where α overwhelms the kinetic rates, and smaller values of α only degrade the performance of the algorithm. It should be noted that, because of non-Markovian effects present in real molecular systems, it may not be possible to achieve the large α limit in molecular RE simulations.

In Figure 2.6B we show the effect of α on the number of transition events per replica for two replicas as a function of the high temperature T_2. As expected, the number of transition events becomes smaller as α decreases. The drop in the number of events is most dramatic when α approaches the magnitude of the conformational transition rate constants (10–100 μs^{-1}). If we compare $N_{\mathrm{TE}}(\tau|T_1)/N$ with the expected number of transitions for a single-temperature simulation at T_1 (Figure 2.6B dashed line), we see that for some combinations of α and T_2 the efficiency of two-replica NRE is less than a uncoupled non-RE simulation, while for others the efficiency is improved.

The value of T_2 which maximizes the number of transition events also decreases as α decreases. This arises due to a competition between the increase in the number of transition events at high temperature as T_2 approaches 440 K (the temperature at which the harmonic mean rate is maximized) and the decrease in the efficiency in transfer of those transitions to the low temperature by temperature exchanges due to the decrease of w with increasing temperature gap. Thus, there is a temperature for which there is an optimal balance between the increasing number of conformational transition events at high temperature and the decreasing efficiency of transfer to low temperature. This optimum occurs when the two competing effects are of comparable magnitude, leading to a decrease in the optimum temperature as α decreases.

The finite-α behavior of NRE for many replicas is more complex, as issues related to

the size of the state space become important. While in the limit of infinite α, any conformational transition in a replica at any temperature is "communicated" via rapid temperature exchanges to T_1 before the replica has had a chance to move back, this is not the case for finite α. The most apparent symptom of this is that a simulation with more replicas can be less efficient than one with fewer. This can be seen in Figure 2.5B, where the insertion of additional replicas into a fixed temperature range can lead to a decrease in $N_{\mathrm{TE}}(\tau|T_1)/N$. This is related to the rapid increase in the combinatoric size of the NRE state space as N increases. As defined previously, a transition event is counted only when the system evolves from a state where the replica of interest is U at T_1 to one in which it is F (also at T_1) and back. For example, when the system leaves a composite state of the form $U_1XX\ldots X$, it must find its way to a state of the form $F_1XX\ldots X$ and back for a transition event to occur for replica 1. However, the number of states of the form $F_1XX\ldots X$ for N replicas is $2^{N-1}(N-1)!$, and the ratio of the size of this "target set" of states to the total number of accessible states $2^N N!$ decreases as N^{-1} when N increases. The more replicas there are in the NRE simulation, the longer any excursion in temperature space away from T_1 will last, and we expect the number of transition events to reflect this.

2.3.4 Convergence efficiency of the kinetic network under special conditions

In order to study the effect of increasing the number of replicas on the efficiency of NRE in isolation, we studied the NRE model for the case in which the folding and unfolding rate constants are independent of temperature. The effect of different temperature distributions and changing temperature exchange rates with different numbers of replicas are thereby excluded, and the efficiency of NRE will only be affected by the size of the combinatorial state space. This temperature-independent system is equivalent to one in which all of the replicas are starting at the same temperature, but where each temperature is distinguishable

by a virtual label. We define a transition event as before, i.e. a round-trip change in conformational state of a replica conditional on a temperature label. We will refer to this special temperature label as "the temperature of interest".

Since all replicas are equivalent in terms of their kinetic properties and there is no increased rate of conformational interconversion at high temperature that the replica at the temperature of interest can "borrow from", we expect that the number of transitions per replica will at best match that of a single-processor simulation. Specifically, we expect that the number of transition events per replica to be small for small α, and that it will increase monotonically as a function of α, approaching the number of transitions for a single-processor simulation as $\alpha \to \infty$. If we examine the behavior of the total number of transition events at the temperature of interest as a function of α for various numbers of replicas N (Figure 2.7), we see that this is indeed the case. However, the value of α needed to give a value for the number of transition events close to the asymptotic limit depends strongly on N: for $N = 2$, $\alpha \approx 100\ \mu\mathrm{s}^{-1}$ is sufficiently large to approximate the infinite limit, while $\alpha \approx 10\ \mathrm{ns}^{-1}$ is required when $N = 10$. For $N = 40$, even larger values of α are required. This is a direct consequence of the increase in the combinatoric complexity of the search space, in that increasingly larger temperature interconversion rates are required to propagate a conformational change at a distant temperature to the temperature of interest in a time that is short compared to the conformational transition rates.

This increase in combinatoric complexity is also seen in the behavior of the number of transitions per replica as the number of replicas increases (Figure 2.8). This is similar to the effect seen in Figure 3 of the main paper, however here all possible contributions of the temperature dependence of the rate constants have been eliminated. Again, we see a strong decrease in the efficiency as the number of replicas increases, reflecting the increased possibility of a replica becoming "lost" in the combinatoric state space.

The origin of these phenomena originate fundamentally from the increase in the size of the state space, and consequently, from the increase in the average time a given replica

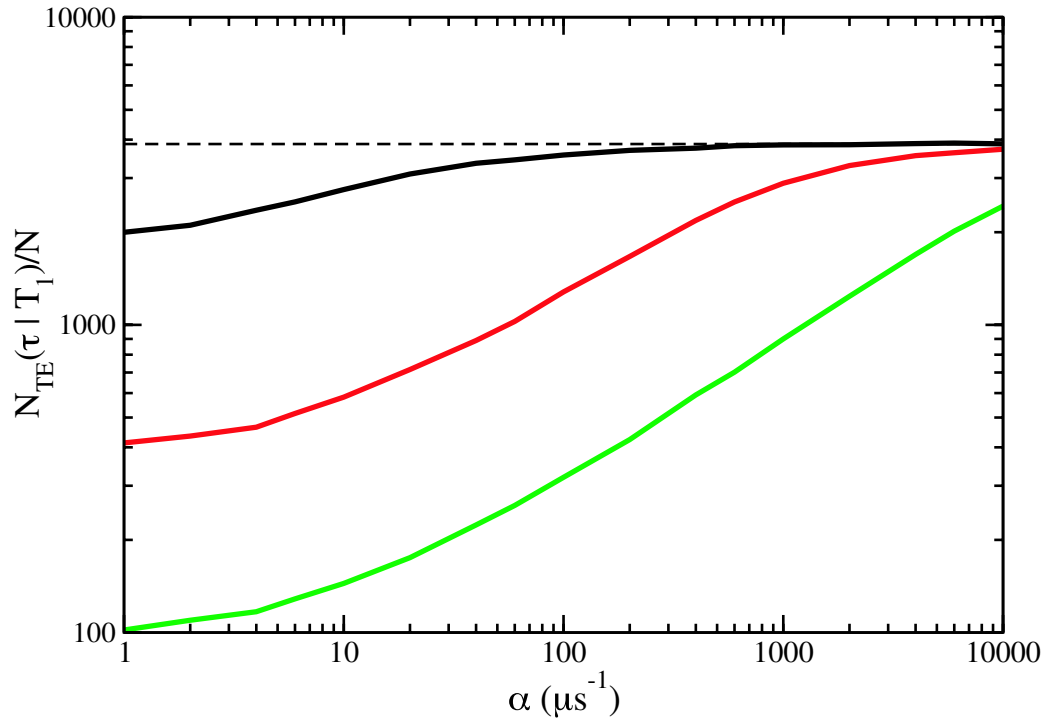

Figure 2.7: Number of transition events per replica as a function of α for a temperature-independent rate system in a total simulation time of 4 ms. The folding and unfolding rates were those of the anti-Arrhenius model at 440 K (i.e. $k_u = 12.06\ \mu s^{-1}$ and $k_f = 1.052\ \mu s^{-1}$). The predicted number of transition events for an uncoupled, non-RE simulation with the same rates and simulation time is shown as a black dashed line and corresponds to the $\alpha \to \infty$ limit. The black, red and green data correspond to $N = 2$, 10, and 40, respectively.

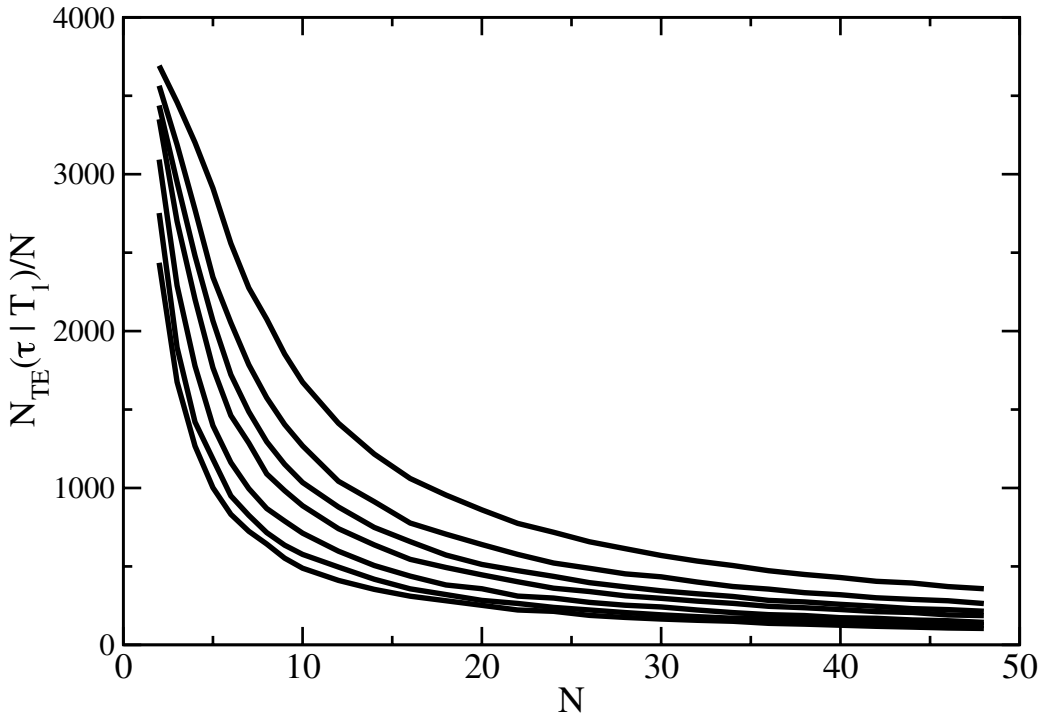

Figure 2.8: Number of transition events per replica as a function of the number of replicas N for a temperature-independent rate system in a total simulation time of 4 ms. The folding and unfolding rates were those of the anti-Arrhenius model at 440 K (i.e. $k_u = 12.06\ \mu\mathrm{s}^{-1}$ and $k_f = 1.052\ \mu\mathrm{s}^{-1}$). The curves correspond to $\alpha = 5$, 10, 20, 40, 60, 100, and 200 $\mu\mathrm{s}^{-1}$ from bottom to top, respectively.

spends in a single excursion away from the temperature of interest. Since for a given replica there are on the order of N states in which that replica is not at the temperature of interest for every state in which that replica is at the temperature of interest, on average, the time a given replica spends in a single excursion away from the temperature of interest increases linearly with N (Figure 2.9).

It should be noted that although the efficiency is degraded when α is small and the number of replicas is large, the correct fraction folded at low temperature can nonetheless be obtained with the NRE model for anti-Arrhenius folding rates for as many as 20 replicas (data not shown). This is despite the fact that for 20 replicas there are $2^{20}\ 20! \approx 10^{24}$ composite states, and therefore it is not possible for any NRE simulation to visit each state once, much less reach equilibrium. This demonstrates that it is possible to achieve convergence of average quantities without the convergence of the full replica exchange ensemble. This is not unreasonable, since the convergence of any one of the $N!$ symmetry-related sub-networks is sufficient to obtain correct macrostate populations. Therefore, even a local exploration of the full kinetic network is sufficient to obtain converged results.

2.4 Conclusions

In this chapter we have used a kinetic network model of replica exchange to explore the effects of anti-Arrhenius behavior of the conformational kinetics on the convergence of replica exchange protein folding simulations. We have constructed a network model for replica exchange inspired by protein folding and have studied its convergence behavior as a function of the number of replicas, their temperatures, the kinetics at each temperature, and the rate of temperature exchange. The number of folding transitions is used as an indicator for convergence. The results demonstrate that the convergence of NRE for a two replica system in the limit of very rapid temperature exchanges is fastest when the high temperature is chosen to maximize the harmonic mean of the folding and unfolding rates.

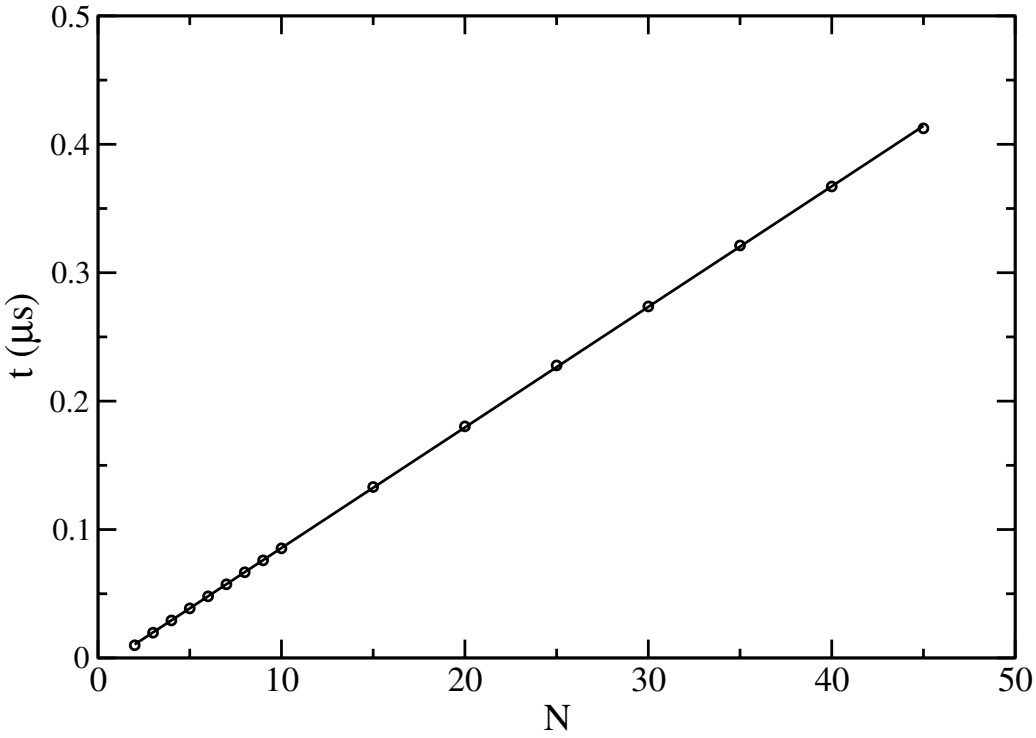

Figure 2.9: The average amount of time t spent in a given excursion in temperature space away from the temperature of interest for a temperature-independent rate system as a function of the number of replicas N. The folding and unfolding rates were those of the anti-Arrhenius model at 440 K (i.e. $k_u = 12.06\ \mu\mathrm{s}^{-1}$ and $k_f = 1.052\ \mu\mathrm{s}^{-1}$) and $\alpha = 100\ \mu\mathrm{s}^{-1}$. The line is a least-squares fit.

Additional replicas improve the efficiency in the NRE model only if the harmonic mean of the kinetic rates at the temperature of the additional replica is larger than the average of the harmonic means of the original set of replicas. Both the convergence rate and efficiency are reduced if the temperature exchange rate is finite, and the optimal temperature of the high-temperature is reduced.

The conclusions obtained here are based on the behavior of a simplified network model of replica exchange which is completely Markovian. More of the characteristics of molecular RE could be incorporated into the NRE model to enhance its realism. For example, continuous energy distributions could be used to simulate the effects of energy distribution overlaps. Non-Markovian effects, such as non-exponential waiting time distributions could also be modeled, either directly or by dividing the F and U macrostates into "hidden" microstates. Even though many proteins are observed to follow simple two state kinetics for folding under some conditions, the underlying free energy landscape is undoubtedly more complex. The NRE model can also be extended to simulate more complex landscapes represented by three or many more macrostates. It could turn out that the best strategies for optimizing RE simulations are different for such cases as compared with those where the kinetics is described by two state anti-Arrhenius behavior as has been observed for some peptides[35, 38].

The results shown here for the NRE model are nevertheless likely to be relevant for atomic-level RE simulations, and suggest that more extensive "training" simulations to explore the temperature dependence of the kinetics will be useful for optimizing the efficiency of RE. Training simulations have been used to construct asynchronous variants of RE[88] and to find the optimum temperature ladder by maximizing the diffusion in temperature space[16, 29]. However, maximizing the diffusion of replicas in temperature space regardless of the actual kinetics at each temperature does not necessarily optimize the RE simulation. If the rate constants have anti-Arrhenius behavior, then there exists an optimal temperature with the fastest kinetics. Additional replicas beyond that temperature decrease

the efficiency of the simulation relative to the case where the same number of replicas were used, but where the additional replicas are placed close to the optimum temperature. This is because in the anti-Arrhenius case the optimum temperature has more favorable kinetic properties than any higher temperature, and can contribute more to the convergence of the low temperature of interest. In this context, finding the optimum high temperature should take priority, and the remaining replicas can then be distributed to optimize temperature diffusion and efficiency. On the other hand, in the context of Arrhenius-like rates, there is no optimum high temperature, and the focus on the optimization of diffusion to the highest temperature is justified.

The possibility that an arbitrary choice of highest temperature may be too high is further increased by the observation that finite temperature exchange rates lower the optimal highest temperature significantly below that predicted by the harmonic mean of the forward and reverse rates at high temperature. Superficially, it could be argued that this result is not relevant to atomic-level simulations, which are already conducted in the "large-α" limit, given that the folding and unfolding timescales of peptides and small proteins are on the order of tens to hundreds of nanoseconds while temperature exchanges are typically done on a picosecond timescale. However, unlike the NRE model, for which temperature exchanges of any magnitude can freely occur, in a molecular simulation the rate of temperature exchanges is limited by the rate of diffusion in energy space. For example, a replica must first find low-energy configurations to be able to exchange temperature with a replica at a lower temperature. Therefore, the rate of conformational transitions places an upper limit on the effective value of α that can be achieved in a molecular simulation.

NRE also provides some insights into the choice of the number of replicas and their temperature distribution. In molecular RE simulations, the temperature spacing is dictated primarily by the overlap of energy distributions at different temperatures. However, if we wish to add additional replicas beyond those required to obtain sufficient energy overlap (for example, in a large-scale cluster or grid computing environment), the NRE results

indicate that additional replicas will be most beneficial to efficiency if they are placed at temperatures such that the average of the harmonic means is increased. Additionally, it may be possible to use re-weighting methods such as T-WHAM[89], which generate estimates of thermodynamic quantities based on data from more than one temperature, to further accelerate convergence properties, since folding transitions are not required to occur between identical temperatures to be "productive". RE methods which are based on the exchange of energy function parameters[90] may also have more favorable convergence properties for some systems.

The replica exchange technique is a powerful conformational sampling method for the study of quasi-ergodic systems while preserving canonical thermodynamic properties. For these reasons, it has become a very popular tool in computational biophysics research. This study identifies some characteristics of the method that are key for the effective use of RE to study processes with anti-Arrhenius kinetic behavior, such as protein folding and binding.

2.5 Appendix I: Closed form analysis for the $\alpha \to \infty$ limit of the network model

In the large α limit, the network model can be greatly simplified and some neat analytical treatment can be done to get meaningful results that can also be verified via simulations. We begin with the full cubic model of Figure 2.2. As α becomes large, the "up-and-down" transitions become very fast relative to the transitions along the top and bottom faces of the cube. Therefore, we can assume that the pairs of states connected by the vertical transitions (corresponding to temperature exchanges) experience instantaneous equilibration, and can be considered as single states. This reduces the number of effective states from 8 to 4: $FF \equiv \{F_1F_2, F_2F_1\}$, $FU \equiv \{F_1U_2, F_2U_1\}$, $UF \equiv \{U_1F_2, U_2F_1\}$, and $UU \equiv \{U_1U_2, U_2U_1\}$. We can imagine that each of these 4 composite states has "inside of it" the two temperature-labeled states with their respective equilibrium probabilities. For

example, the state FF has "inside" F_1F_2 and F_2F_1 each with population $1/2$ (since they have equal populations at equilibrium), while FU has "inside" F_1U_2 and F_2U_1 with populations $1/(1+w)$ and $w/(1+w)$, respectively. In terms of kinetics, the rate to exit a given state is simply the population-weighted sum of the rates corresponding to the "internal" sub-states. For example, the rate corresponding to the $FU \to UU$ transition will be the rate for $F_1U_2 \to U_1U_2$ weighted by $1/(1+w)$ (the relative population of F_1U_2 "inside" FU) plus the rate for $F_2U_1 \to U_2U_1$ weighted by $w/(1+w)$ (the relative population of F_2U_1 "inside" FU). Working through these sums, we end up with the square network of Figure 2.10 with rates

$$k_A = \frac{1}{2}(k_{u1} + k_{u2})$$

$$k_B = \frac{k_{f1}k_{f2}(k_{u1} + k_{u2})}{k_{f1}k_{u2} + k_{u1}k_{f2}}$$

$$k_C = \frac{k_{u1}k_{u2}(k_{f1} + k_{f2})}{k_{f1}k_{u2} + k_{u1}k_{f2}}$$

and

$$k_D = \frac{1}{2}(k_{f1} + k_{f2}).$$

The kinetic matrix for the network in Figure 2.10 has three non-zero eigenvalues, given by $\lambda_1 = k_B + k_C$ and

$$\lambda_\pm = \frac{1}{2}\left[\eta \pm \sqrt{\eta^2 - 8(k_Ak_C + 2k_Ak_D + k_Bk_D)}\right],$$

where $\eta = 2k_A + k_B + k_C + 2k_D$. I have shown that these rates give the correct equilibrium probabilities (proof omitted) and have numerically confirmed that the eigenvalues are the same as those obtained for the 8×8 kinetic matrix corresponding to the full network with a very large value of α.

The distribution of temperature-unconditional first passage times is related to the kinetics of the network where the destination states have been replaced by a single absorbing state (Figure 2.11). The two non-zero eigenvalues of the corresponding kinetic matrix are

$$\lambda_\pm = \frac{1}{2}(2k_A + k_B + k_C \pm \rho),$$

where $\rho = \sqrt{4k_A(k_A - k_C) + (k_B + k_C)^2}$. These eigenvalues are clearly distinct, and will in general lead to a bi-exponential first passage time distribution. Solving for the first passage time distribution (by solving the master equation for $P(U,t)$ and differentiating with respect to t) we find that

$$P(t_{FP}) = \frac{1}{2\rho}[2k_A(1-p^0_{FF})+k_C(2p^0_{FF}-1)+k_B](\lambda_- e^{-\lambda_- t}-\lambda_+ e^{-\lambda_+ t})+\frac{1}{2}(\lambda_- e^{-\lambda_- t}+\lambda_+ e^{-\lambda_+ t}),$$

where the initial populations are $P(FF,0) = p^0_{FF}$ and $P(FU,0) = 1 - p^0_{FF}$. The mean first passage time can be obtained analytically by integration:

$$\begin{aligned}\langle t_{FP}\rangle &= \int_0^\infty t_{FP} P(t_{FP})\, dt_{FP} \\ &= \frac{2k_A + k_B + p^0_{FF}(k_C - k_A)}{k_A k_B + 2k_A k_C}.\end{aligned}$$

For the specific case of the rates used in our simulations ($k_{f1} = 0.818$, $k_{u1} = 0.13$, $k_{f2} = 1.05$, $k_{u2} = 12.06$), the collapsed rates are $k_A = 6.095$, $k_B = 1.0468$, and $k_C = 0.2928$, and the non-zero eigenvalues are $\lambda_+ = 12.7492$ and $\lambda_- = 0.7804$, with corresponding eigenvectors $e_+ = (-2.14685, 1.14685, 1)$ and $e_- = (-0.08404, -0.91596, 1)$. Since the equilibrium populations $(0,0,1)$ minus e_- is $(0.08404, 0.91596, 0)$, choosing initial populations in which $p^0_{FF} = 0.08404$ will lead to a single exponential first passage time distribution with rate $\lambda_- = 0.7804$. On the other hand, because of the pattern of signs in e_+, it is impossible to find initial conditions for which the first passage time distribution is a single exponential with rate λ_+. The mean first passage time is 1.2449, close in magnitude to $1/\lambda_- = 1.281$, but considerably longer than $1/\lambda_+ = 0.078$. Our simulations results confirmed that the two eigenvalues λ_+ and λ_- match with the double exponential curve-fit parameters from the plot of the mean first passage time.

The eigenvalue or master equation approach inherently cannot give information about the paths taken to reach equilibrium. In the simple system of Figure 2.11, however, it is possible to make some analytical statements about the paths. Suppose, for example, that we begin from state FF at time 0. There are three types of paths by which FF can

reach U: $FF \to U$ (1-step path of type 1), $[FF \to FU]_n \to U$ ($2n$-step path of type 2, $n = 1, 2, \ldots$), and $[FF \to FU]_n \to FF \to U$ ($2n+1$-step path of type 3, $n = 1, 2, \ldots$). The probability of the type 1 path occurring is $k_A/(k_A + k_A) = 1/2$, while the probabilities of paths of type 2 and 3 are

$$P_2(n) = \frac{k_B^{n-1} k_C}{2^n (k_B + k_C)^n}$$

and

$$P_3(n) = \frac{k_B^n}{2^{n+1} (k_B + k_C)^n}$$

respectively. These probabilities are normalized, since

$$\begin{aligned} \sum_{n=1}^{\infty} P_2(n) &= \frac{k_C}{2(k_B + k_C)} \sum_{n=0}^{\infty} \left(\frac{k_B}{2(k_B + k_C)} \right)^n \\ &= \frac{k_C}{2(k_B + k_C)} \left(1 - \frac{k_B}{2(k_B + k_C)} \right)^{-1} \\ &= \frac{k_C}{2k_C + k_B}, \end{aligned}$$

$$\begin{aligned} \sum_{n=1}^{\infty} P_3(n) &= \frac{1}{2} \sum_{n=1}^{\infty} \left(\frac{k_B}{2(k_B + k_C)} \right)^n \\ &= \frac{1}{2} \left[\left(1 - \frac{k_B}{2(k_B + k_C)} \right)^{-1} - 1 \right] \\ &= \frac{k_B}{2(2k_C + k_B)}, \end{aligned}$$

and

$$\frac{1}{2} + \frac{k_C}{2k_C + k_B} + \frac{k_B}{2(2k_C + k_B)} = 1.$$

If we use the rates of our simulation, the fraction of paths starting at FF that are absorbed via type 1, 2, and 3 paths are 0.5, 0.18, and 0.32, respectively. Thus, approximately 68% of the paths are paths in which the final transition to reach U is from FF, nearly $3/4$'s of which occur directly as the first transition after time $t = 0$. (Thanks for Dr. Andrec providing me this derivation.)

2.6 Appendix II: Publication attached

Part of the contents of this chapter was published in *Proc. Natl. Acad. Sci. USA*, 104, 15340-15345 (2007).

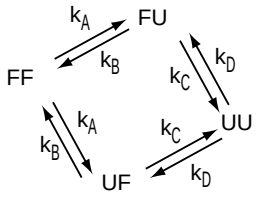

Figure 2.10: The collapsed 4-state kinetic network model.

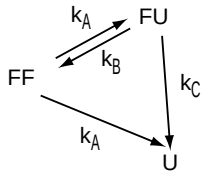

Figure 2.11: The collapsed kinetic network model with an absorbing state corresponding to walker 1 unfolded.

Simulating replica exchange simulations of protein folding with a kinetic network model

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, NJ 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, NJ 08854

Edited by David Baker, University of Washington, Seattle, WA, and approved August 1, 2007 (received for review May 18, 2007)

Replica exchange (RE) is a generalized ensemble simulation method for accelerating the exploration of free-energy landscapes, which define many challenging problems in computational biophysics, including protein folding and binding. Although temperature RE (T-RE) is a parallel simulation technique whose implementation is relatively straightforward, kinetics and the approach to equilibrium in the T-RE ensemble are very complicated; there is much to learn about how to best employ T-RE to protein folding and binding problems. We have constructed a kinetic network model for RE studies of protein folding and used this reduced model to carry out ''simulations of simulations'' to analyze how the underlying temperature dependence of the conformational kinetics and the basic parameters of RE (e.g., the number of replicas, the RE rate, and the temperature spacing) all interact to affect the number of folding transitions observed. When protein folding follows anti-Arrhenius kinetics, we observe a speed limit for the number of folding transitions observed at the low temperature of interest, which depends on the maximum of the harmonic mean of the folding and unfolding transition rates at high temperature. The results shown here for the network RE model suggest ways to improve atomic-level RE simulations such as the use of ''training'' simulations to explore some aspects of the temperature dependence for folding of the atomic-level models before performing RE studies.

anti-Arrhenius | Markov process | parallel tempering

One of the key challenges in the computer simulation of proteins at the atomic level is the sampling of conformational space. The efficiency of many common sampling protocols, such as Monte Carlo (MC) and molecular dynamics (MD), is limited by the need to cross high free-energy barriers between conformational states and rugged energy landscapes. One class of methods for studying equilibrium properties of quasi-ergodic systems that has received a great deal of recent attention is based on the replica exchange (RE) algorithm (1, 2) (also known as parallel tempering). To accomplish barrier crossings, RE methods simulate a series of replicas over a range of temperatures. Periodically, coordinates are exchanged by using a Metropolis criterion (3) that ensures that at any given temperature a canonical distribution is realized. RE methods, particularly REMD (4), have become very popular for the study of protein biophysics, including peptide and protein folding (5, 6), aggregation (7–9), and protein–ligand interactions (10, 11). Previous studies of protein folding appear to show a significant increase in the number of reversible folding events in REMD simulations versus conventional MD (12, 13). Given the wide use of REMD, a better understanding of the RE algorithm and how it can be used most effectively for the study of protein folding and binding is of considerable interest.

The effectiveness of RE methods is determined by the number of temperatures (replicas) that are simulated, their range and spacing, the rate at which exchanges are attempted, and the kinetics of the system at each temperature. Although the determination of ''optimal'' Metropolis acceptance rates and temperature spacings has been the subject of various studies (2, 14–19), the role played by the intrinsic temperature-dependent conformational kinetics that is central to understanding RE has not received much attention. Recent work (19–22) recognizes the importance of exploration of conformational space and the crossing of barriers between conformational states as the key limiting factor for the RE algorithm. Molecular kinetics can have a strong effect on RE beyond the entropic effects that have been discussed (20, 22), particularly if the kinetics does not have simple temperature dependence. It is known from experimental and computational studies that the folding rates of proteins and peptides can exhibit anti-Arrhenius behavior, where the folding rate decreases with increasing temperature (23–28). Different models have been proposed to explain the physical origin of this effect (29, 30).

In this paper, we investigate the impact of simulation parameters and anti-Arrhenius kinetics on the RE method. Because RE simulations of protein systems that display anti-Arrhenius behavior are difficult to converge, we developed a network RE (NRE) model that allows us to simulate the RE algorithm of two-state protein folding. This network model reduces the atomic complexity of the system to a set of discrete conformational states that evolve in continuous time according to Markovian kinetics for both conformational transitions and exchange between replicas.

The NRE model studied here does not capture all of the complexities of the ''real'' molecular simulation because various kinds of non-Markovian behavior are not captured in the network model. However, it does capture some of the essential features of RE and allows us to study these fundamental aspects of the algorithm in a controlled setting and at low computational cost, which allows us to separate some of the interacting parameters and study their effects on the simulation individually. Many of the limitations in the convergence rates and efficiency observed with NRE also will be present in full atomic-level RE simulations, allowing us to identify promising avenues of inquiry for future atomic-level simulations.

Theory

The RE Method and the NRE Model. In a standard RE simulation with M replicas corresponding to M inverse temperatures $\beta_i = (k_B T_i)^{-1}$ ($\beta_1 > \beta_2 > \ldots > \beta_M$), the state of the extended ensemble is specified by a joint configuration of M replicas $X = \{x_1, x_2, \ldots, x_M\}$, where x_i stands for the configuration of replica i. To simulate the extended ensemble, a propagation algorithm such as MC or constant-temperature MD is used to locally

Author contributions: M.A., E.G., and R.M.L. designed research; W.Z. and M.A. performed research; W.Z. analyzed data; and W.Z., M.A., E.G., and R.M.L. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Abbreviations: RE, replica exchange; T-RE, temperature RE; MC, Monte Carlo; MD, molecular dynamics; NRE, network RE.

¶To whom correspondence should be addressed. E-mail: ronlevy@lutece.rutgers.edu.

sample the conformational space within each replica, and exchanges of configurations between pairs of replicas, e.g., $X = \{\ldots, x_i, \ldots, x_j, \ldots\} \rightarrow X' = \{\ldots, x_j, \ldots, x_i, \ldots\}$ are attempted periodically with an acceptance probability $w(X \rightarrow X')$. For the equilibrium distribution to remain invariant with respect to these exchanges, it is sufficient to impose a detailed balance condition on the transition probability. For the potential energy function $U(x)$, the appropriate transition probability is given by (4)

$$w(X \rightarrow X') = \min\{1, \exp[-(\beta_j - \beta_i)(U(x_i) - U(x_j))]\}. \quad [1]$$

To isolate some of the essential features of the RE algorithm, we construct a kinetic NRE model, which we can use to study the effects of the parameters of the model on efficiency and convergence. We consider a system in which the configurational space can be partitioned into two macrostates of interest separated by a free-energy barrier that makes transitions between the conformations an activated process. Motivated by protein folding, we call these macrostates F and U (for "folded" and "unfolded"). Transitions between F and U in a (non-RE) MD or kinetic MC simulation can be approximated by a Poisson process in which the waiting times between folding and unfolding transition events are exponentially distributed random variables with means equal to the reciprocal of the folding or unfolding rates, respectively.

If the transition events are Markovian, then we can represent the simultaneous behavior of two noninteracting replicas in terms of the four composite states $\{F_1F_2, F_1U_2, U_1F_2, U_1U_2\}$. In each symbol, the first letter is the configuration of replica 1, the second letter is the configuration of replica 2, and the subscripts are the temperatures of each replica. Therefore, F_1U_2 represents the composite state that replica 1 at temperature T_1 is folded, while replica 2 at temperature T_2 is unfolded. The kinetics in the composite state space can be represented as a continuous-time Markov process with discrete states (31).

The four-state composite system corresponding to noninteracting replicas can be extended to create a discrete-state model of RE by introducing temperature exchanges between replicas. For example, suppose the current state is F_1U_2. After a successful temperature exchange, replica 1 is at T_2 and replica 2 is at T_1, thus the new state can be represented as F_2U_1. The introduction of temperature exchange therefore creates four additional states, leading to the eight-state system $\{F_1F_2, F_1U_2, U_1F_2, U_1U_2, F_2F_1, F_2U_1, U_2F_1, U_2U_1\}$. These states are arranged into two subnetworks defined by the "horizontal" folding and unfolding transitions, which are connected to each other by "vertical" temperature-exchange transitions, forming a cubic network (Fig. 1). In general, the network for an N-replica system consists of $N!$ subnetworks, each of which has 2^N states connected by folding/unfolding transitions. The model description in this section will focus primarily on the two-replica case; all of the details can be generalized easily to the case of N replicas.

We require that the equilibrium populations of the states be such that the canonical ensemble is recovered at each temperature. This is the case if the equilibrium populations are proportional to the product of the equilibrium populations for the two-state systems, e.g.,

$$P_{eq}(F_1U_2) = \frac{1}{2} P_{eq}(F_1)P_{eq}(U_2) = \frac{1}{2} \frac{k_{f1}k_{u2}}{(k_{f1} + k_{u1})(k_{f2} + k_{u2})},$$

where the factor of 1/2 accounts for the presence of the two equivalent manifolds. For these probabilities to be preserved under temperature exchanges, it is sufficient that detailed balance is satisfied, e.g., the transition probabilities $w(F_1U_2 \rightarrow F_2U_1)$ and $w(F_2U_1 \rightarrow F_1U_2)$ satisfy $P_{eq}(F_1U_2)w(F_1U_2 \rightarrow F_2U_1) = P_{eq}(F_2U_1)w(F_2U_1 \rightarrow F_1U_2)$ or

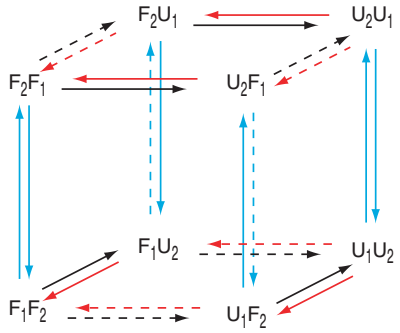

Fig. 1. The kinetic network of the composite states corresponding to the simplified RE model with two replicas. The state labels represent the conformation (letter) and temperature (subscript) for each replica. For example, F_2U_1 represents the state in which replica 1 is folded and at temperature T_2 while replica 2 is unfolded and at temperature T_1. Red and black arrows correspond to folding and unfolding transitions, respectively, and the temperature at which the transition occurs is indicated by the solid and dashed lines (for T_2 and T_1, respectively). The cyan arrows correspond to temperature-exchange transitions, with the solid and dashed cyan lines denoting transitions with rate parameters α and $w\alpha$, respectively.

$$\frac{w(F_1U_2 \rightarrow F_2U_1)}{w(F_2U_1 \rightarrow F_1U_2)} = \frac{k_{f2}k_{u1}}{k_{f1}k_{u2}} \equiv w. \quad [2]$$

If the equilibrium favors the folded state at T_1 and the unfolded state at T_2, then $w < 1$. The ratios of forward and reverse transition probabilities for $F_1F_2 \rightleftharpoons F_2F_1$ and $U_1U_2 \rightleftharpoons U_2U_1$ are equal to one because interchange of temperatures does not change the equilibrium populations.

In atomic-level RE simulations, temperature-exchange attempts usually are made periodically in time, i.e., the MC or MD evolution is interrupted, temperature swap proposal(s) are made, and the proposals are either accepted or rejected (4, 6). In keeping with the continuous-time nature of our network model, we simulate the effect of temperature exchanges by introducing an additional rate parameter α, which controls the overall scaling of the temperature-exchange rate relative to the folding and unfolding rates. We set the forward and reverse rates of the $F_1F_2 \rightleftharpoons F_2F_1$ and $U_1U_2 \rightleftharpoons U_2U_1$ "reactions" equal to α, while the other rates are set to α or $w\alpha$ (Fig. 1) as required by detailed balance (Eq. **2**), and where we choose $w < 1$. For example, the states U_1F_2 and U_2F_1 differ in population, with U_2F_1 being more populated if the equilibrium favors the folded state at T_1 and the unfolded state at T_2. We therefore set the $U_1F_2 \rightarrow U_2F_1$ "reaction rate" equal to α and the reverse rate equal to $w\alpha$, where w is defined in Eq. **2**.

The NRE model can be simulated by using a standard method for continuous-time Markov processes with discrete states (31), also known as the "Gillespie algorithm." The algorithm remains efficient even when the number of replicas is large (e.g., 20 replicas, corresponding to 10^{24} states) because of the fact that each state is connected to a small number of neighboring states (those connected by single temperature exchanges involving neighboring temperatures and folding/unfolding transitions of each replica).

The convergence or efficiency of a simulation is monitored by measuring $N_{TE}(\tau|T_1)$, the number of "round-trip" transitions between the U and F states, conditional on the temperature of interest T_1 that occurs in a given observation time τ. In the context of the network model, suppose that we follow replica 1, and at a given time the system is in a state where that replica is folded at temperature T_1 (e.g., F_1F_2). We then wait for the first occurrence of a state in which replica 1 is unfolded at T_1 (e.g., U_1F_2) and then for the first occurrence of a state in which that

replica is folded again at T_1 (e.g., F_1F_2). At this point, we say that a transition event has occurred. Conceptually, a transition event is a transit of a given replica from one conformation at low temperature to the other conformation at low temperature and back again regardless of route, i.e., whether it was the result of a direct barrier crossing at T_1 or indirectly via a barrier crossing at T_2 combined with temperature exchanges. The number of transitions as defined corresponds to the number of "reversible folding" events studied in all-atom simulations of peptide systems (12, 13).

Thermodynamic Model for Anti-Arrhenius Behavior. The Arrhenius equation relates a reaction rate k to the temperature:

$$k(T) = Ae^{-\Delta G^{\dagger}(T)/k_BT} = Ae^{-(\Delta E^{\dagger}(T) - T\Delta S^{\dagger}(T))/k_BT}, \quad [3]$$

where $\Delta G^{\dagger}(T)$ is the free energy of activation. The temperature dependence of the reaction rate customarily is described by means of the Arrhenius plot, the plot of $\ln k(T)$ with respect to $1/T$. The slope of $\ln k(T)$ in the Arrhenius plot is proportional to the activation energy, $\Delta E^{\dagger}(T)$, at temperature T. When the activation energy is temperature-independent, the Arrhenius plot appears as a line of constant slope. Moreover, if the activation energy is positive, the reaction rate increases with increasing temperature. This behavior is referred to as normal Arrhenius behavior. When the activation energy is negative, however, increasing the temperature causes the rate to decrease. This nonintuitive phenomenon sometimes observed in protein folding kinetics (23–28) is referred to as anti-Arrhenius behavior. In these circumstances, the transition state is energetically favored but entropically disfavored with respect to the reactants.

Often protein folding rates follow normal Arrhenius behavior at low temperatures, switching to anti-Arrhenius behavior at higher temperatures. This mixed behavior can be understood in terms of a constant activation heat-capacity model in which the activation energy and entropy vary linearly with respect to the temperature and its logarithm, respectively (24, 32):

$$\Delta E^{\dagger}(T) = \Delta E^{\dagger}(T_0) + (T - T_0)\Delta C_p^{\dagger} \quad [4]$$

$$\Delta S^{\dagger}(T) = \Delta S^{\dagger}(T_0) + \ln(T/T_0)\Delta C_p^{\dagger}, \quad [5]$$

where $\Delta C_p^{\dagger} < 0$ is the activation heat capacity, which is assumed here to be independent of temperature. Summing Eqs. **4** and **5**, we obtain the expression for $\Delta G^{\dagger}(T)$ corresponding to this model. Shown in Fig. 2 are the Arrhenius plots for the unfolding and folding rates, $k_u(T)$ and $k_f(T)$, used in this work that result from inserting this expression in Eq. **3**, setting $\ln A/s^{-1} = 22$, $T_0 =$ 300 K, and $\Delta E^{\dagger}(T_0)$, $\Delta S^{\dagger}(T_0)$, and $\Delta C_p^{\dagger}$ to be 2 kcal/mol, −0.01 kcal/mol·K, and −0.025 kcal/mol·K for folding, and 8.5 kcal/mol, 0.008 kcal/mol·K, and 0 kcal/mol·K for unfolding, respectively. For the case of Arrhenius folding (Fig. 2, dashed line), the parameters are identical with the exception that $\Delta C_p^{\dagger}$ for folding is zero. The unfolding rate follows normal linear Arrhenius behavior, whereas the anti-Arrhenius folding rate decreases with increasing temperature above $T^* = 380$ K (the temperature at which the activation energy for folding is zero and the folding rate is maximal). The general behavior of $k_u(T)$ and $k_f(T)$ shown in Fig. 2 is typical for experimentally determined peptide folding kinetic rates (23, 25, 28).

Results

We have measured the number of conformational transitions for the NRE model by using both Arrhenius and anti-Arrhenius models for the folding and unfolding rates with various choices of the number of replicas, their temperatures, and the temperature-exchange rate parameter α. The goal of these calculations is to study factors that affect the increased efficiency that RE can provide. We define the efficiency in the context of NRE to be the total number of transitions divided by the number of replicas $N_{TE}(\tau|T_1)/N$. We make several general observations. First, increasing the total temperature range for a given number of replicas can degrade the efficiency of reversible folding if the kinetics is anti-Arrhenius (Fig. 3*A*). To understand this behavior, we first examine the behavior of NRE for the simple case of two replicas ($N = 2$), where the rate of temperature exchanges is large compared with the folding/unfolding kinetics. The condition that α be very large relative to the conformational kinetic rates simplifies the problem because in that limit the behavior is

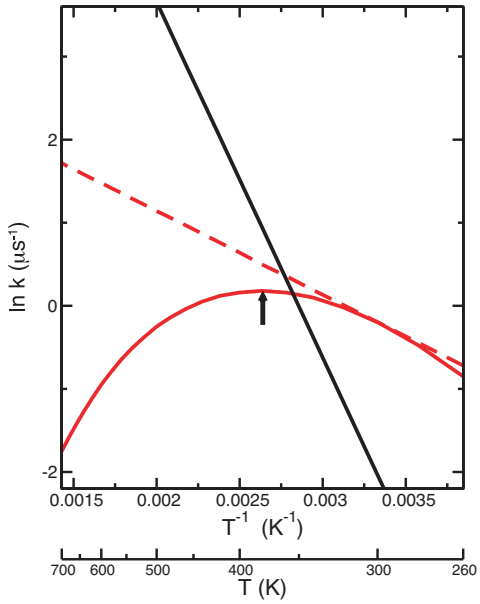

Fig. 2. Arrhenius plot of the folding and unfolding rates from a thermodynamic model for the temperature dependence of protein folding rate constants. The black line corresponds to the unfolding rate, and the red lines correspond to the folding rates. The solid line is for the $\Delta C_p^{\ddagger} \neq 0$ case displaying anti-Arrhenius behavior, whereas the dashed line corresponds to the same parameters with $\Delta C_p^{\ddagger} = 0$. The arrow indicates the temperature T^* at which the folding rate is maximal (≈380 K).

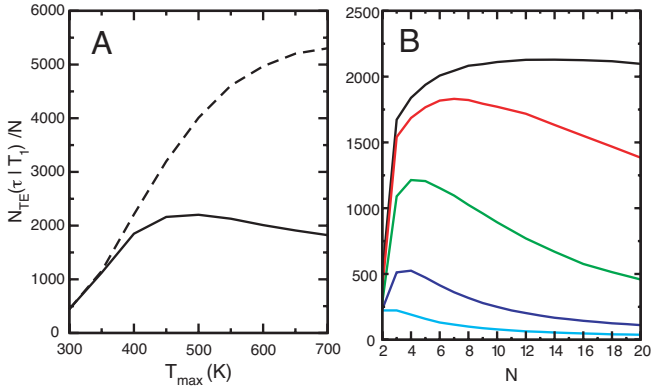

Fig. 3. Number of transition events in NRE simulations (normalized by the number of replicas) for various temperature ranges, exchange rates α, and number of replicas N. In all cases, the system was simulated for $\tau = 4$ ms. For the simulations in *A*, α was set to 1,000 μs^{-1}, the dashed and solid lines correspond to Arrhenius and anti-Arrhenius kinetics, respectively, and six replicas were exponentially distributed between 300 K and T_{max}. The simulations in *B* were performed with anti-Arrhenius rates, N replicas exponentially distributed from 300 K to 700 K, and α values of 10,000 μs^{-1} (black), 1,000 μs^{-1} (red), 100 μs^{-1} (green), 10 μs^{-1} (blue), and 1 μs^{-1} (cyan).

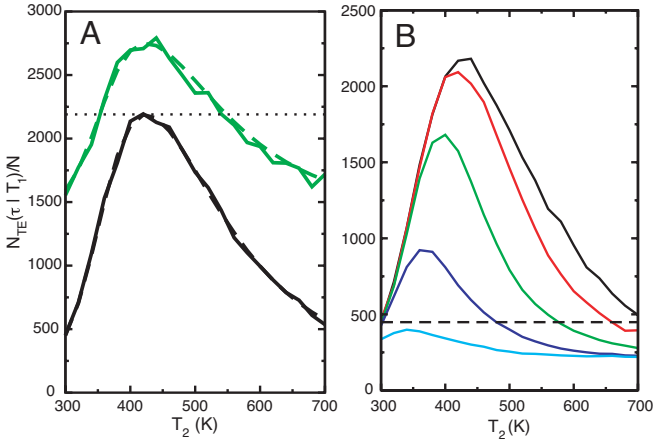

Fig. 4. Number of transition events per replica in NRE simulations using the anti-Arrhenius folding rates for a simulation time $\tau = 4$ ms conditional on temperature $T_1 = 300$ K, while T_2 is scanned from 300 K to 700 K. (*A*) Solid black and green lines show simulation results for two-replica and three-replica systems (with $T_3 = 440$ K), respectively. Dashed black and green lines show the number of transition events predicted by using the average of harmonic means for two and three replicas, respectively. All simulations were performed with $\alpha = 10$ ns^{-1}. (*B*) Results for two-replica NRE simulations using the anti-Arrhenius folding rates and α values of 10 ns^{-1} (solid black), 1 ns^{-1} (red), 100 μs^{-1} (green), 10 μs^{-1} (blue), and 1 μs^{-1} (cyan). The dashed black line corresponds to the predicted number of transitions for a single uncoupled simulation at T_1.

independent of the precise choice of α and depends on the (temperature-dependent) folding and unfolding rates. We fix T_1 at 300 K and sweep T_2 over the range 300 K to 700 K. In Fig. 4*A*, we show the dependence of $N_{TE}(\tau|T_1)$ normalized by the number of replicas as a function of T_2 for the anti-Arrhenius kinetic model. We see that $N_{TE}(\tau|T_1)/N$ is small at low and high T_2 and reaches a maximum near 440 K (Fig. 4*A*, solid black line).

The number of transition events at the low temperature T_1 obtained by simulation in the large α limit is very well approximated by the average of the harmonic means of the folding and unfolding rates at both temperatures:

$$N_{TE}(\tau|T_1)/N \approx \frac{\tau}{N}[(k_{f1}^{-1} + k_{u1}^{-1})^{-1} + (k_{f2}^{-1} + k_{u2}^{-1})^{-1}] \qquad [6]$$

(Fig. 4*A*, dashed black line). For the uncoupled, non-RE case, the rate of transition events at each temperature is simply the harmonic mean of the rate constants. Therefore, our observation (Eq. **6**) suggests that the number of transition events observed at the lowest temperature in the coupled RE case can be no larger than the number of transitions at an "optimum" temperature defined as that temperature for which the number of folding/unfolding transitions for the uncoupled system is maximized. Because the number of transitions for the uncoupled system is a harmonic mean of the rate constants, the overall convergence of NRE at low temperature is limited by the smallest rate at this optimum (higher) temperature.

Next, we examine how the number of replicas affects the convergence as monitored by the number of transition events. In Fig. 4*A*, we examine whether a third replica results in an improvement over the optimum behavior with two replicas by fixing T_1 at 300 K and T_3 at 440 K (the two-replica optimum) and scanning T_2 from 300 K to 700 K (i.e., we do not require $T_1 < T_2 < T_3$). We see in Fig. 4*A* (solid green line) that the number of transitions per replica again reaches a maximum near $T_2 \approx 440$ K, corresponding to the case in which one replica is at the temperature of interest (300 K) and the other two are both placed at the "optimal" temperature of 440 K. As in the two-replica case, $N_{TE}(\tau|T_1)/N$ is very well approximated by the average of the harmonic means of the rates at all three temperatures (Fig. 4*A*, dashed green line).

The relevant question is whether the addition of the third replica is an improvement over having two. It is important in this regard to distinguish the convergence rate from the computational efficiency of the simulation. In the cases seen in Fig. 4*A*, the total number of transition events (not normalized by the number of walkers) is larger for three replicas than the maximum total number of transition events for two replicas, and therefore we expect the convergence to be better. In general, adding an additional replica always will improve overall convergence, because the additional transition pathways opened up always will have a positive contribution to the total number of transition events. However, the computational efficiency of NRE as measured by $N_{TE}(\tau|T_1)/N$ of the three-replica simulation is improved relative to the two-replica simulation only if the additional temperature T_2 has values between 350 K and 550 K (Fig. 4*A*, dotted black line). Although the addition of a replica always improves convergence, it improves efficiency only if the harmonic mean of the rates at the additional temperature is large relative to the harmonic means of the other replicas. If not, then the presence of the additional slow paths will reduce the efficiency. For the general case of NRE with N replicas, we expect that, in the large α limit, optimal efficiency (and convergence) will be obtained when one replica is at the temperature of interest and all of the other replicas are placed at the temperature that maximizes the harmonic mean of the folding and unfolding rates. Thus, the replica with the largest harmonic mean sets a "speed limit" for the amount of efficiency improvement that an RE simulation can have over an uncoupled simulation run for the same amount of CPU time. The addition of replica $N + 1$ will increase the efficiency only if the harmonic mean at the new temperature is greater than the average of the harmonic means of the original N replicas.

In the results described above, the rate of temperature exchanges is so large that convergence is limited only by the rates of conformational transitions at each temperature. When α is comparable to or smaller than the rates of conformational transitions, the waiting time for a temperature exchange to occur becomes comparable to or even larger than the timescale of configuration changes within each replica. Therefore, there can be multiple folding or unfolding events at higher temperatures before any of these events are transmitted to the temperature of interest. These events are "lost" and make no contribution to the number of transition events at low temperature. Therefore, in the NRE model (where conformational transitions are instantaneous and strictly Markovian), the optimal convergence (and efficiency) is achieved in the limit where α overwhelms the kinetic rates, and smaller values of α only degrade the performance of the algorithm. It should be noted that, because of non-Markovian effects present in real molecular systems, it may not be possible to achieve the large α limit in molecular RE simulations.

In Fig. 4*B*, we show the effect of α on the number of transition events per replica for two replicas as a function of the high temperature T_2. As expected, the number of transition events becomes smaller as α decreases. The drop in the number of events is most dramatic when α approaches the magnitude of the conformational transition rate constants (10–100 μs^{-1}). If we compare $N_{TE}(\tau|T_1)/N$ with the expected number of transitions for a single-temperature simulation at T_1 (Fig. 4*B*, dashed line), we see that for some combinations of α and T_2 the efficiency of two-replica NRE is less than an uncoupled non-RE simulation, whereas for others the efficiency is improved.

The value of T_2 that maximizes the number of transition events also decreases as α decreases. This result arises because of a competition between the increase in the number of transition

events at high temperature as T_2 approaches 440 K (the temperature at which the harmonic mean rate is maximized) and the decrease in the efficiency in transfer of those transitions to the low temperature by temperature exchanges caused by the decrease of w with increasing temperature gap. Thus, there is a temperature for which there is an optimal balance between the increasing number of conformational transition events at high temperature and the decreasing efficiency of transfer to low temperature. This optimum occurs when the two competing effects are of comparable magnitude, leading to a decrease in the optimum temperature as α decreases.

The finite-α behavior of NRE for many replicas is more complex because issues related to the size of the state space become important. Although in the limit of infinite α, any conformational transition in a replica at any temperature is "communicated" via rapid temperature exchanges to T_1 before the replica has had a chance to move back, this is not the case for finite α. The most apparent symptom of this is that a simulation with more replicas can be less efficient than one with fewer, which can be seen in Fig. 3*B*, where the insertion of additional replicas into a fixed temperature range can lead to a decrease in $N_{TE}(\tau|T_1)/N$. This result is related to the rapid increase in the combinatoric size of the NRE state space as N increases.

Conclusions

In this paper, we have used a kinetic NRE model to explore the effects of anti-Arrhenius behavior of the conformational kinetics on the convergence of RE protein folding simulations. We have constructed a NRE model inspired by protein folding and have studied its convergence behavior as a function of the number of replicas, their temperatures, the kinetics at each temperature, and the rate of temperature exchange. The number of folding transitions is used as an indicator for convergence. The results demonstrate that the convergence of NRE for a two-replica system in the limit of very rapid temperature exchanges is fastest when the high temperature is chosen to maximize the harmonic mean of the folding and unfolding rates. Additional replicas improve the efficiency in the NRE model only if the harmonic mean of the kinetic rates at the temperature of the additional replica is larger than the average of the harmonic means of the original set of replicas. Both the convergence rate and efficiency are reduced if the temperature-exchange rate is finite, and the optimal temperature of the high temperature is reduced.

The conclusions obtained here are based on the behavior of a simplified NRE model, which is completely Markovian. More of the characteristics of molecular RE could be incorporated into the NRE model to enhance its realism. For example, continuous energy distributions could be used to simulate the effects of energy-distribution overlaps. Non-Markovian effects, such as nonexponential waiting time distributions also could be modeled, either directly or by dividing the F and U macrostates into "hidden" microstates. Even though many proteins are observed to follow simple two-state kinetics for folding under some conditions, the underlying free-energy landscape is undoubtedly more complex. The NRE model also can be extended to simulate more complex landscapes represented by three or many more macrostates. It could turn out that the best strategies for optimizing RE simulations are different for such cases as compared with those in which the kinetics is described by two-state anti-Arrhenius behavior as has been observed for some peptides (25, 28).

The results shown here for the NRE model nevertheless are likely to be relevant for atomic-level RE simulations, and they suggest that more extensive "training" simulations to explore the temperature dependence of the kinetics will be useful for optimizing the efficiency of RE. Training simulations have been used to construct asynchronous variants of RE (33) and to find the optimum temperature ladder by maximizing the diffusion in temperature space (6, 19). However, maximizing the diffusion of replicas in temperature space regardless of the actual kinetics at each temperature does not necessarily optimize the RE simulation. If the rate constants have anti-Arrhenius behavior, then there exists an optimal temperature with the fastest kinetics. Additional replicas beyond that temperature decrease the efficiency of the simulation relative to the case in which the same number of replicas are used but the additional replicas are placed close to the optimum temperature. The reason for this is because in the anti-Arrhenius case the optimum temperature has more favorable kinetic properties than any higher temperature and can contribute more to the convergence of the low temperature of interest. In this context, finding the optimum high temperature should take priority, and the remaining replicas then can be distributed to optimize temperature diffusion and efficiency. On the other hand, in the context of Arrhenius-like rates, there is no optimum high temperature, and the focus on the optimization of diffusion to the highest temperature is justified.

The possibility that an arbitrary choice of highest temperature may be too high is increased further by the observation that finite temperature-exchange rates lower the optimal highest temperature significantly below that predicted by the harmonic mean of the forward and reverse rates at high temperature. Superficially, it could be argued that this result is not relevant to atomic-level simulations, which already are conducted in the "large-α" limit, given that the folding and unfolding timescales of peptides and small proteins are on the order of tens to hundreds of nanoseconds, whereas temperature exchanges typically are done on a picosecond timescale. However, unlike the NRE model, for which temperature exchanges of any magnitude can occur freely, in a molecular simulation the rate of temperature exchanges is limited by the rate of diffusion in energy space. For example, a replica must first find low-energy configurations to be able to exchange temperature with a replica at a lower temperature. Therefore, the rate of conformational transitions places an upper limit on the effective value of α that can be achieved in a molecular simulation.

NRE also provides some insights into the choice of the number of replicas and their temperature distribution. In molecular RE simulations, the temperature spacing is dictated primarily by the overlap of energy distributions at different temperatures. However, if we wish to add additional replicas beyond those required to obtain sufficient energy overlap (for example, in a large-scale cluster or grid computing environment), the NRE results indicate that additional replicas will be most beneficial to efficiency if they are placed at temperatures such that the average of the harmonic means is increased. Additionally, it may be possible to use reweighting methods such as T-WHAM (34), which generate estimates of thermodynamic quantities based on data from more than one temperature, to further accelerate convergence properties because folding transitions are not required to occur between identical temperatures to be "productive." RE methods that are based on the exchange of energy function parameters (35) also may have more favorable convergence properties for some systems.

The RE technique is a powerful conformational sampling method for the study of quasi-ergodic systems while preserving canonical thermodynamic properties. For these reasons, it has become a very popular tool in computational biophysics research. This study identifies some characteristics of the method that are key for the effective use of RE to study processes with anti-Arrhenius kinetic behavior, such as protein folding and binding.

We thank Attila Szabo for helpful discussions. This work was supported in part by National Institutes of Health Grant GM 30580.

1. Swendsen RH, Wang J-S (1986) *Phys Rev Lett* 57:2607–2609.
2. Hukushima K, Nemoto K (1996) *J Phys Soc Jpn* 65:1604–1608.
3. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) *J Chem Phys* 21:1087–1091.
4. Sugita Y, Okamoto Y (1999) *Chem Phys Lett* 314:141–151.
5. Rhee YM, Pande VS (2003) *Biophys J* 84:775–786.
6. Nymeyer H, Gnanakaran S, García AE (2004) *Methods Enzymol* 383:119–149.
7. Cecchini M, Rao F, Seeber M, Caflisch A (2004) *J Chem Phys* 121:10748–10756.
8. Tsai H-HG, Reches M, Tsai C-J, Gunasekaran K, Gazit E, Nussinov R (2005) *Proc Natl Acad Sci USA* 102:8174–8179.
9. Baumketner A, Shea J-E (2005) *Biophys J* 89:1493–1503.
10. Verkhivker GM, Rejto PA, Bouzida D, Arthurs S, Colson AB, Freer ST, Gehlhaar DK, Larson V, Luty BA, Marrone T, Rose PW (2001) *Chem Phys Lett* 337:181–189.
11. Ravindranathan KP, Gallicchio E, Friesner RA, McDermott AE, Levy RM (2006) *J Am Chem Soc* 128:5786–5791.
12. Rao F, Caflisch A (2003) *J Chem Phys* 119:4035–4042.
13. Seibert MM, Patriksson A, Hess B, van der Spoel D (2005) *J Mol Biol* 354:173–183.
14. Kofke DA (2002) *J Chem Phys* 117:6911–6914.
15. Kone A, Kofke DA (2005) *J Chem Phys* 122:206101.
16. Predescu C, Predescu M, Ciobanu CV (2004) *J Chem Phys* 120:4119–4128.
17. Predescu C, Predescu M, Ciobanu CV (2005) *J Phys Chem B* 109:4189–4196.
18. Rathore N, Chopra M, de Pablo JJ (2005) *J Chem Phys* 122:024111.
19. Trebst S, Troyer M, Hansmann UHE (2006) *J Chem Phys* 124:174903.
20. Zuckerman DM, Lyman E (2006) *J Chem Theory Comput* 2:1200–1202.
21. Zuckerman DM, Lyman E (2006) *J Chem Theory Comp* 2:1693.
22. Beck DAC, White GWN, Daggett V (2007) *J Struct Biol* 157:514–523.
23. Segawa S-I, Sugihara M (1984) *Biopolymers* 23:2473–2488.
24. Oliveberg M, Tan Y-J, Fersht AR (1995) *Proc Natl Acad Sci USA* 92:8926–8929.
25. Muñoz V, Thompson PA, Hofrichter J, Eaton WA (1997) *Nature* 390:196–199.
26. Karplus M (2000) *J Phys Chem B* 104:11–27.
27. Ferrara P, Apostolakis J, Caflisch A (2000) *J Phys Chem B* 104:5000–5010.
28. Yang WY, Gruebele M (2004) *Biochemistry* 43:13018–13025.
29. Scalley ML, Baker D (1997) *Proc Natl Acad Sci USA* 94:10636–10640.
30. Bryngelson JD, Wolynes PG (1989) *J Phys Chem* 93:6902–6915.
31. Gillespie DT (1992) *Markov Processes: An Introduction for Physical Scientists* (Academic, Boston).
32. McQuarrie DA, Simon JD (1997) *Physical Chemistry: A Molecular Approach* (University Science Books, Sausalito, CA).
33. Hagen M, Kim B, Liu P, Friesner RA, Berne BJ (2007) *J Phys Chem B* 111:1416–1423.
34. Gallicchio E, Andrec M, Felts AK, Levy RM (2005) *J Phys Chem B* 109:6722–6731.
35. Liu P, Kim B, Friesner RA, Berne BJ (2005) *Proc Natl Acad Sci USA* 102:13749–13754.

Chapter 3

Simulating Replica Exchange simulation of Protein Folding with a continuous two-dimensional potential model

3.1 Introduction

Replica exchange (RE) methods [11, 91, 12, 92, 14] are widely employed to enhance the conformational sampling efficiency of biomolecular simulations for the study of protein biophysics, including peptide and protein folding[15, 16] and aggregation[17, 18, 19], and protein-ligand interactions[20, 21]. To accomplish barrier crossings, RE methods simulate a series of replicas over a range of potential parameters[93, 94, 95, 96, 97] or temperatures[14]. In the latter, replicas exchange temperatures following a Metropolis criterion designed to preserve canonical distributions. This scheme allows conformations at physiological temperatures, where conformational interconversions are rare, to switch to higher temperatures where transitions to other conformations are more likely. In a sense, therefore, the enhancement of conformational sampling at low temperatures is achieved by "borrowing" the faster kinetics at higher temperatures.

The popularity of RE methods is due to their ease of implementation and their ability to enhance conformational sampling while preserving canonical distributions at the thermodynamic conditions of each replica. The properties of the RE algorithm and how it can be utilized most effectively for the study of protein folding and binding has received attention recently[30, 32, 98]. The determination of the temperature assignment and number of replicas to achieve optimal temperature mixing has been the subject of a variety of

studies[12, 24, 25, 26, 27, 28, 29, 99]. Recent work has also recognized the importance of conformational relaxation as a key limiting factor which can affect the efficiency of the RE algorithm [30, 31, 29, 32]. While temperature RE is relatively straightforward to implement, kinetics in the RE ensemble is complicated and does not correspond in any simple way to the molecular kinetics (necessitating additional methods for the reconstruction of molecular kinetics from RE samples[79, 80, 100, 73]). Molecular kinetics, however, can have a strong effect on RE, especially when the kinetics has complex temperature dependence. The anti-Arrhenius behavior typical of protein folding kinetics, where the folding rate above a critical threshold temperature decreases with increasing temperature [34, 36, 37, 38], is understood to occur when the transition state is energetically favored but entropically disfavored with respect to the reactants. Anti-Arrhenius behavior represents a challenge for temperature RE because when folding exhibits anti-Arrhenius behavior there exists a temperature (generally unknown) at which the folding and unfolding rates are optimal. If even higher temperatures beyond the optimal are included in the RE ensemble, this may degrade performance[98].

Although some comparative studies aimed at determining the benefits of RE over conventional MD for peptide folding have been conducted[101, 32, 102], it is far from straightforward to systematically explore the convergence properties of RE by brute force molecular simulations, since RE simulations of protein folding are very difficult to converge. To understand some of the basic mechanisms that determine the efficiency of RE it is useful to study simplified low dimensionality systems that share some of the key characteristics of molecular systems. In chapter2 we investigated a discrete two-state network model for replica exchange (NRE), containing two conformational states (Folded and Unfolded) at each of several temperatures[98]. We found that the efficiency of RE for this system varies non-monotonically with respect to the temperature distribution of the replicas when the folding rate displays anti-Arrhenius behavior. The model showed that the rate of folding/unfolding events in RE is maximal when high temperature replicas are placed near the

temperature at which the harmonic mean of the folding and unfolding rates for the uncoupled system (k_f and k_u) is maximal. This result suggested that in molecular simulations adding high temperature replicas does not necessarily lead to increased efficiency of exploration of conformational space, and that, instead, optimal efficiency could be obtained by placing replicas at specific temperatures determined by the temperature dependence of key kinetic rates of the system.

In this chapter we extend this analysis by studying a continuous two-dimensional system designed to reproduce the anti-Arrhenius kinetics of a conformational equilibrium, such as a protein folding equilibrium, mediated by an entropic bottleneck. The two-dimensional system studied here is an extension of the potential model we originally used to study the convergence of the weighted histogram analysis method,[89] and is very similar in spirit to to the funnel-like golf course model for protein folding studied by Szabo and co-workers[103]. This two-dimensional system is sufficiently simple to be amenable to accurate analytical and numerical solution, while including some characteristics of molecular systems that were absent from the discrete NRE model. The present model is self-contained in that the kinetic rates are determined by the potential and the move set rather than being imposed, as in the NRE model of reference [98]. Furthermore, and most importantly, the unfolded and folded macrostates have, like real molecular systems, microscopic internal structure. The new model makes it possible to follow the joint microscopic evolution of the system in conformational and temperature space. It incorporates the same discrete temperature exchange scheme commonly adopted in replica exchange molecular simulations, and it allows us to study the effects of non-Markovian processes likely present in replica exchange simulations of molecular systems.

In the next section we present the potential model and the kinetic scheme we have employed. We review the replica exchange method and the network model for replica exchange we previously developed. We then summarize the thermodynamic and kinetic properties of the two-dimensional system and present results showing how these determine

the efficiency of the replica exchange method. This chapter is then concluded by discussing the implications of these findings for replica exchange simulations of molecular systems.

3.2 Methods

3.2.1 The two-dimensional continuous potential

A two-dimensional potential was constructed to mimic the anti-Arrhenius temperature dependence of the folding rate seen in proteins. We designed this potential to have an energetic barrier when going from the "folded" to the "unfolded" 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 3.1):

$$B(x) = \begin{cases} \delta, & -1 \leq x \leq 0 \\ bx^{n_1} + \delta, & 0 < x \leq 1 \end{cases} . \tag{3.1}$$

The use of a boundary of this form is based on a two-dimensional potential first used in our laboratory to study the convergence of the weighted histogram analysis method[89], and is very similar in spirit to simplified models for protein folding studied by Bicout and Szabo[103] and the model of an entropic barrier by Zhou and Zwanzig [104]. The specific parameters δ, b, and n_1 were chosen together with the parameters of the potential function discussed below by trial and error to achieve a sufficiently strong temperature dependence to illustrate some of the possible consequences of anti-Arrhenius behavior on RE simulations. It is natural to choose the x axis to be the reaction coordinate, with $-1 \leq x \leq 0$ corresponding to the folded macrostate and $0 < x \leq 1$ to the unfolded macrostate. The move set was chosen to be compatible with this reaction coordinate (see below). In order for folding and unfolding to be activated processes, however, it is necessary to add a potential energy function which has an energetic well as a function of x in the folded region,

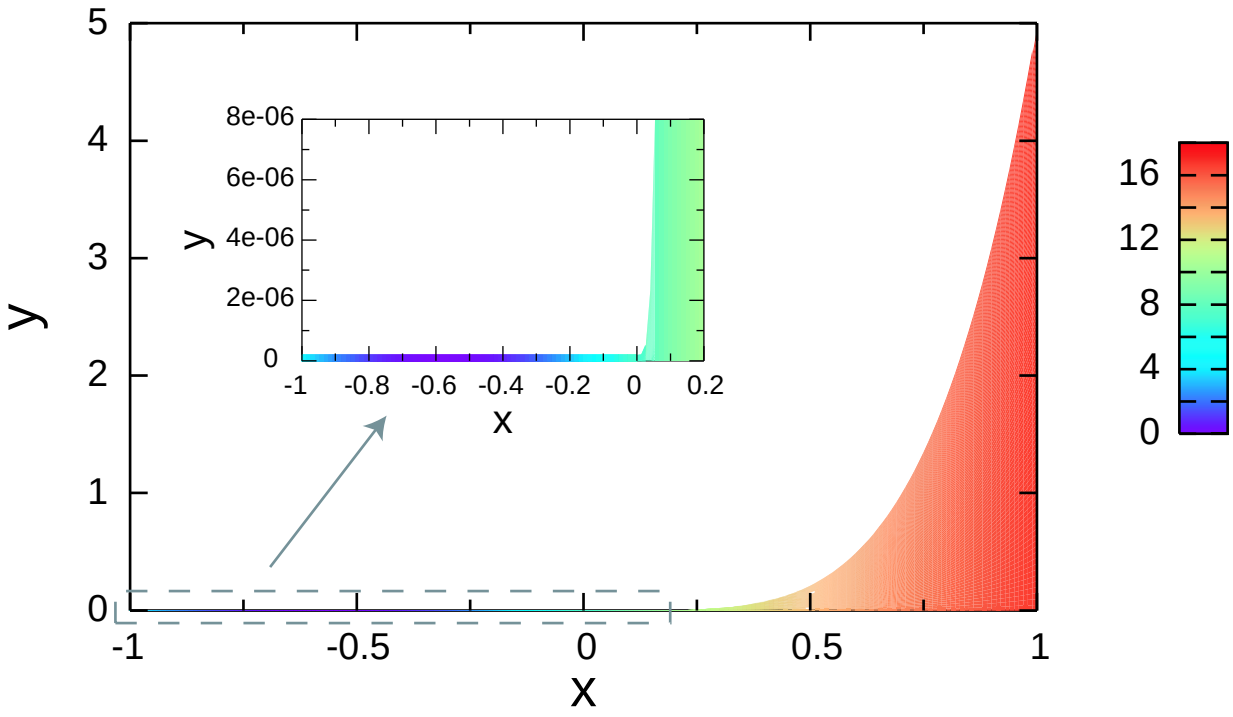

Figure 3.1: A schematic representation of the two-dimensional potential function used in this work. The colored area corresponds to the accessible region of the (x, y) plane, with the colors representing the magnitude of the potential energy at that (x, y) point (scale bar in kcal/mol). The potential energy is infinite in the non-colored region and for $y < 0$, $x < -1$, and $x > 1$. The inset is an enlarged view of the folded macrostate and transition region.

and increases with x in the unfolded region. Specifically, we use the potential function

$$U(x,y) = \begin{cases} a_1(x+x_0)^2, & -1 \le x < -x_1 \quad 0 \le y \le B(x) \\ -a_2x^2 + c_0, & -x_1 \le x \le 0, \quad 0 \le y \le B(x) \\ a_3x^{n_2} + c_0, & 0 < x < x_2, \quad 0 \le y \le B(x) \\ a_4x^{n_3} + c_1 & x_2 \le x \le 1, \quad 0 \le y \le B(x) \\ \infty & \text{otherwise} \end{cases}, \tag{3.2}$$

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, $c_0 = 7.059$ kcal/mol, $b = 5$, $n_1 = 4.55$, $n_2 = 2$, $n_3 = 0.5$, and $\delta = 2 \times 10^{-7}$. The constants $x_0 = \sqrt{c_0(a_1+a_2)/a_1a_2}$, $x_1 = a_1x_0/(a_1+a_2)$, $x_2 = (a_4n_3/a_3n_2)^{\frac{1}{n_2-n_3}}$, $c_1 = c_0 - (a_4x_2^{n_3} - a_3x_2^{n_2})$ were chosen so that the first derivative of $U(x,y)$ is continuous. A graphical representation of the two-dimensional system studied here is shown in Figure 3.1.

3.2.2 Kinetics on the two-dimensional continuous potential

We use Metropolis Monte Carlo (MC) sampling to simulate the movement of a particle in this two-dimensional potential. Kinetic MC has a long history in the study of protein folding using simplified models[105, 106, 107]. To ensure rapid equilibration along the y coordinate consistent with the choice of x as the reaction coordinate and because of the large size difference of the accessible region in the y direction between the folded and unfolded regions, we adopted an asymmetric MC proposal scheme,[108, 89] in which the step size in the y direction is proportional to $b(x)$, i.e. a proposed move (x', y') is generated uniformly in the region $x - \Delta < x' < x + \Delta$, $y - b(x)\Delta < y' < y + b(x)\Delta$. The displacement parameter Δ was chosen such that the barrier crossing is slow but not prohibitively expensive and follows a linear regime (i.e. doubling Δ causes approximately a doubling in the number of barrier crossings). To correct for the asymmetric MC proposal distribution, the factor $\theta(|y' - y| / b(x')\Delta)$ was included to satisfy detailed balance, where $\theta(z)$ equals 1 if $z < 1$ and 0 otherwise.

Rate constants in units of MC steps were obtained via MC simulation by calculating the mean first passage times between the two macrostates. The same displacement parameter $\Delta = 0.05$ was used for all temperatures. A "buffer region" $-0.1 < x < 0.0437$ was defined as not belonging to either the folded or unfolded state to reduce artefactual rapid re-crossings of the barrier[109, 110]. For comparison, the temperature dependence of the folding and unfolding rate constants were also estimated from the PMF using the Arrhenius equation $k = A\exp(-\Delta G^{\dagger}/k_B T)$, where $\Delta G^{\dagger}$ is the free energy difference between the transition state and the appropriate macrostate. Free energies were extracted from the PMF along the x axis by averaging the PMF over the macrostates and transition region using numerical integration.

3.2.3 RE simulation on the two-dimensional continuous potential

Replica exchange simulations were performed by running N MC simulations at N inverse temperatures $\beta_i = (k_B T_i)^{-1}$ ($\beta_1 > \beta_2 > \ldots > \beta_N$) in parallel. The state of the extended ensemble is specified by a joint configuration of N replicas $X = \{q_1, q_2, \ldots, q_N\}$, where q_i is the configuration of replica i. Exchanges of configurations were attempted every N_X MC steps between pairs of replicas adjacent in temperature, and the attempted exchange $X = \{\ldots, q_i, q_j, \ldots\} \rightarrow X' = \{\ldots, q_j, q_i, \ldots\}$ was accepted with probability $w(X \rightarrow X')$. Given the potential energy function $U(q)$, the transition probability which satisfies detailed balance and reproduces the canonical ensemble is given by $w(X \rightarrow X') = \min\{1, exp[-(\beta_j - \beta_i)(U(q_i) - U(q_j))]\}$[14].

The efficiency of RE conformational sampling was monitored by measuring $N_{TE}(\tau|T_0)$, the number round-trip transitions in the conformational state of a replica, conditional on the temperature of interest T_0, that occur in a given observation time τ. A transition event is a transit of a given replica from one conformation at T_0 to the other conformation at T_0 *and back again* regardless of route, i.e. whether it was the result of a direct barrier crossing at T_0 or indirectly via a barrier crossing at some other temperature combined with temperature

exchanges. Conceptually, this measure reflects the potential of RE to achieve rapid equilibration at the temperature of interest by means of conformational transitions at temperatures other than the temperature of interest. The transition events as defined correspond to the "reversible folding" events studied in all-atom simulations of peptide systems[22, 23]. We will use the symbol N_{TE} as a shorthand notation for $N_{TE}(\tau|T_0)$, where T_0 will generally be the lowest temperature in the simulation. For an uncoupled simulation, the number of transition events is simply the number of round trips between macrostates.

3.2.4 Review of the discrete Network Replica Exchange (NRE)

We review here the discrete kinetic network model which we devised in our previous study of replica exchange efficiency [98] in chapter2. In this model (unlike the continuous potential model above), the macrostates F and U (for "folded" and "unfolded") do not possess any internal structure. Instead, it is assumed that the system evolves in time as a Poisson process, in which instantaneous transitions between F and U occur after waiting periods given by exponentially distributed random variables with means equal to the reciprocals of the folding or unfolding rates. The result (for a single replica) is an example of a "random telegraph" Markov process[86].

If the transition events are Markovian, then the simultaneous behavior of two uncoupled non-interacting replicas can be represented by the four composite states $\{F_1F_2, F_1U_2, U_1F_2, U_1U_2\}$. In each symbol, the first letter represents the configuration of replica 1, the second letter the configuration of replica 2, and the subscripts denote the temperature of each replica. Only transitions corresponding to a single conformational change (e.g. $F_1F_2 \rightarrow U_1F_2$) are allowed, assuming that the probability of two simultaneous changes (e.g. $F_1U_2 \rightarrow U_1F_2$) in an infinitesimal interval dt can be neglected[86]. The four-state composite system for two non-interacting replicas can be extended to create a network model of replica exchange by introducing temperature exchanges between replicas, i.e. by allowing transitions such as

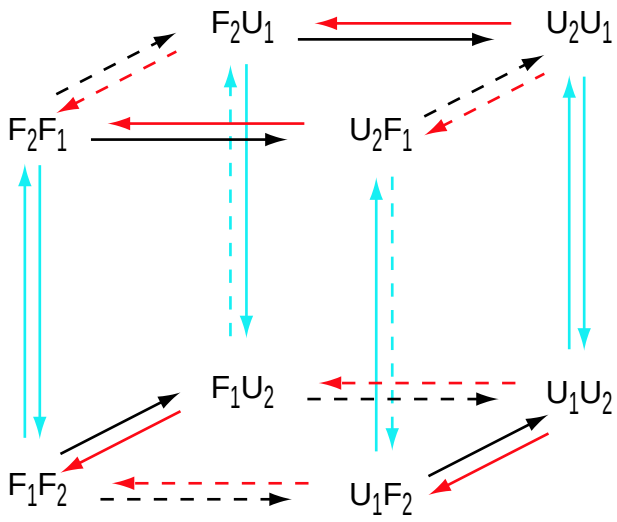

Figure 3.2: The kinetic network model for the discrete NRE model used in chapter 2. The state labels represent the conformation (letter) and temperature (subscript) for each replica. For example, F_2U_1 represents the state in which replica 1 is folded and at temperature T_2, while replica 2 is unfolded and at temperature T_1. Red and black arrows correspond to folding and unfolding transitions, respectively, while the temperature at which the transition occurs is indicated by the solid and dashed lines (for T_2 and T_1, respectively). The cyan arrows correspond to temperature exchange transitions, with the solid and dashed lines denoting transitions with rate parameters α and $w\alpha$, respectively.

$F_1U_2 \rightarrow F_2U_1$. This leads to a system with 8 states arranged in a cubic network with "horizontal" folding and unfolding transitions and "vertical" temperature exchange transitions (Figure 3.2). For canonical equilibrium probabilities to be preserved under temperature exchanges, it is sufficient that detailed balance is satisfied, e.g. the transition probabilities $w(F_1U_2 \rightarrow F_2U_1)$ and $w(F_2U_1 \rightarrow F_1U_2)$ satisfy $P_{eq}(F_1U_2)w(F_1U_2 \rightarrow F_2U_1) = P_{eq}(F_2U_1)w(F_2U_1 \rightarrow F_1U_2)$. The ratios of forward and reverse transition probabilities for $F_1F_2 \rightleftharpoons F_2F_1$ and $U_1U_2 \rightleftharpoons U_2U_1$ are equal to one, as interchange of temperatures does not change the equilibrium populations.

The effect of the rate of temperature exchanges in included by introducing the rate parameter α, which controls the overall scaling of the temperature exchange rate relative to the folding and unfolding rates. The forward and reverse rates of the $F_1F_2 \rightleftharpoons F_2F_1$ and $U_1U_2 \rightleftharpoons U_2U_1$ transitions are set equal to α, while the other rates are set to α or $w\alpha$ as required by detailed balance, where in this case $w = P_{eq}(F_2U_1)/P_{eq}(F_1U_2)$ or its reciprocal such that $w < 1$ (see Figure 3.2). The overall average rate at which temperature exchanges occur (k_{ex}) is the probability of jumping in any instant dt from the upper to the lower face (or *vice versa*) of the cubic network, and is given by the equilibrium population weighted sum of the temperature exchange rates over all states:

$$k_{ex} = \frac{k_{f1}k_{f2} + 2k_{u1}k_{f2} + k_{u1}k_{u2}}{(k_{f1} + k_{u1})(k_{f2} + k_{u2})}\alpha. \tag{3.3}$$

The NRE model was simulated using a standard method for continuous time Markov processes with discrete states[86], also known as the "Gillespie algorithm". Given a current state X_0, we identify its m neighboring states $X_1, X_2, \ldots, X_m$ and the transition rates $k_1, k_2, \ldots, k_m$ from X_0 to each of the neighboring states. We generate a waiting time in state X_0 by drawing a random number from an exponential distribution with mean $(k_1 + k_2 + \cdots + k_m)^{-1}$, and select a destination state X_i from among $X_1, X_2, \ldots, X_m$ with probability $k_i/(k_1 + k_2 + \cdots + k_m)$. This procedure is then repeated with the new state as the current state.

3.3 Results and Discussion

3.3.1 Thermodynamics and kinetics of the continuous model system

a. Thermodynamics

In Figure 3.3 we show the potentials of mean force (PMF) corresponding to the two-dimensional potential along the x coordinate at several temperatures. PMFs calculated by MC sampling and numerical integration of the canonical distribution function agree to within statistical accuracy. The PMFs show two free energy minima corresponding to the folded (F, $x \leq 0$) and unfolded (U, $x > 0$) conformational states, separated by a free energy barrier near $x = 0$. The free energy minimum of the unfolded state and the free energy barrier have no counterparts in the potential, which is monotonically varying in both of these regions (Figure 3.1). These features of the PMF originate from the interplay between opposing entropic and enthalpic driving forces. The free energy minimum of the unfolded state corresponds to the optimal balance between entropy, which drives the system towards large values of x (where the accessible space along the y coordinate is greatest), and enthalpy, which drives the system towards small values of x (where the potential energy is smallest). The free energy barrier that separates the unfolded and folded state is entropic in origin. For x near 0, the entropy is significantly reduced compared to the unfolded state, and assumes a value similar to that of the folded state (compare in Figure 3.1 the size of the accessible space along y at $x = 0$ and for $x > 0$ and $x < 0$). In contrast, the potential energy at $x = 0$, although smaller than in the unfolded state, is still substantially larger than in the folded state. This imbalance between entropy and potential energy causes the free energy maximum at $x = 0$.

From the point of view of folding, the free energy maximum constitutes an entropic bottleneck. In order to make a transition to the folded state, the system needs to cross the free energy barrier region at $x = 0$, where the system has lost all of the entropy required

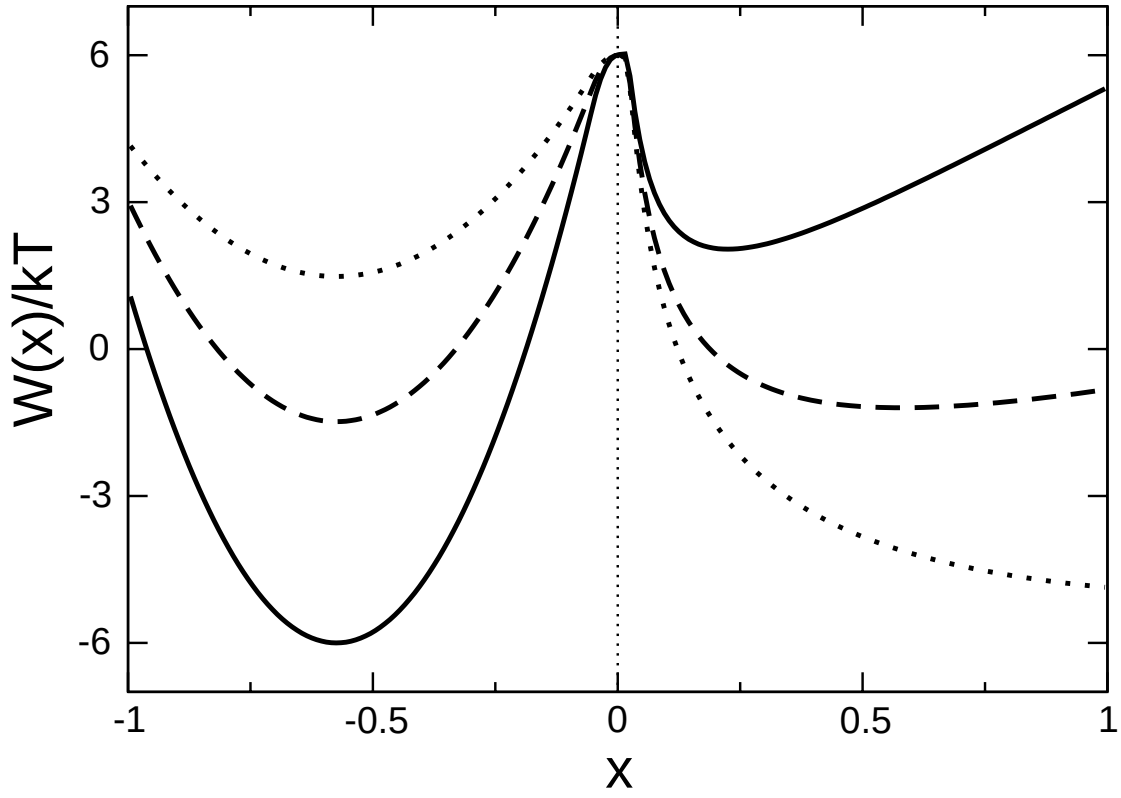

Figure 3.3: The potential of mean force (PMF) at three different temperatures: 296 K (solid), 474 K (dashed) and 789 K (dotted). The PMF was calculated using numerical integration. To more clearly illustrate the change in the barrier height as a function of temperature, the three curves have been superimposed to coincide at $x = 0$.

for folding without having gained all of the folding enthalpy. Similar transition bottlenecks have been described in simplified models for protein folding.[36, 103, 111] After crossing this barrier the system enters the folded state by going downhill in potential energy without further reduction in conformational entropy, since the accessible space along the y direction is the same for all points x in the folded space. Because the conformational entropy is constant for $x < 0$, the potential of mean force in this region coincides with the potential energy. From the point of view of unfolding, the free energy maximum at $x = 0$ constitutes an enthalpic barrier. Relative to the folded state, points in the region near $x = 0$ have similar conformational entropy but larger potential energy. To reach the barrier region from the folded state therefore the system needs to gain potential energy (enthalpy) without the help of a concomitant increase in conformational entropy. Beyond the barrier region there is a free energy gain for moving towards the unfolded state since the gain in conformational entropy outweighs the increase in potential energy.

As shown below, the barrier region close to $x = 0$ constitutes the transition state for the folding/unfolding equilibrium. The free energy difference between the unfolded and folded states and the transition state corresponds to the free energies of activation, which determine the rate of folding and unfolding respectively. Due to their different thermodynamic origins (entropic vs enthalpic), the free energies of activation for folding and unfolding display the opposite dependence on temperature. As Figure 3.3 shows, the free energy of activation for folding increases with increasing temperature relative to thermal energy (kT), where the free energy of activation for unfolding decreases with increasing temperature. This anti-Arrhenius behavior is the signature of an entropically activated process. The conformational entropy difference between the unfolded state and the transition state increases as the temperature is increased, leading to an increase in the height of the free energy barrier for folding with increasing temperature.

Figure 3.4 shows the temperature dependence of the population, $P_F(T)$, of the folded state, often referred to as the melting curve. The shape of the melting curve is typical of

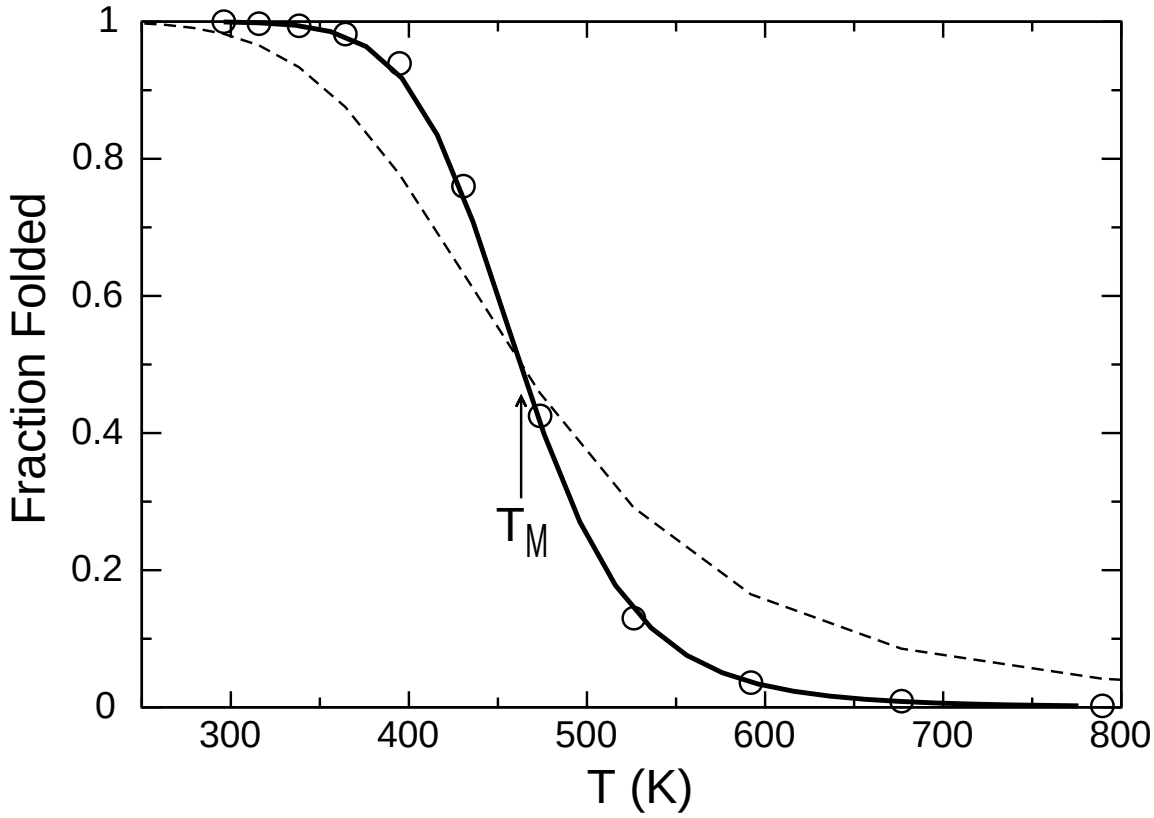

Figure 3.4: The temperature dependence of the fractional population folded (solid line) calculated by numerical integration of the potential of mean force. The temperature dependence of the fraction folded corresponding to a system with a smaller average potential energy difference between the folded and unfolded states (see Appendix I) is shown for comparison (dashed line). The fraction folded derived from the folding and unfolding rates obtained by MC simulation (Figure 3.6) are shown as circles. The melting temperature $T_M = 463$ K (corresponding to 50% folded population) is indicated.

two-state protein thermal denaturation experiments. At $300K$ the system is nearly completely folded, and the fraction folded decreases with increasing temperature in favor of the unfolded state which is entropically favored. The melting temperature T_M (corresponding to equal populations of the folded and unfolded state) is approximately 460 K. At this temperature the folded and unfolded states have equal free energy. The slope of the melting curve at the melting temperature is

$$\left(\frac{dP_F}{dT}\right)_{T=T_M} = \frac{1}{4}\frac{\bar{U}_F - \bar{U}_U}{kT^2},$$

which is proportional to the difference of the average potential energies, U_F and U_U, of the folded and unfolded states. Thus, a decrease of the average potential energy difference between the two states leads to a less steep melting curve. To illustrate this, we show in Figure 3.4 the melting curve corresponding to an alternative parametrization of the potential for which the average potential energy difference between the folded, unfolded, and transition states was decreased, while approximately preserving the same value of the melting temperature (see Appendix I). As expected, the alternative parametrization leads to a more gradual conversion from the folded state to the unfolded state with increasing temperature (Figure 3.4, dashed line). The heat capacity as a function of temperature is approximately Gaussian and is peaked near T_M.

b. Kinetics

With the MC move set described in the Methods Section above, the kinetics of folding/unfolding is two-state as measured by the distribution of first passage times, which is exponential (Figure 3.5). The Arrhenius plots of the folding and unfolding reaction rates are shown in Figure 3.6. The temperature dependence of the reaction rates using the Arrhenius equation with activation free energies extracted from the PMFs (Figure 3.3) agree well with the simulation results, and is a further indication that the kinetics is two-state and that the reaction coordinate is well represented by the x coordinate. This is a consequence of

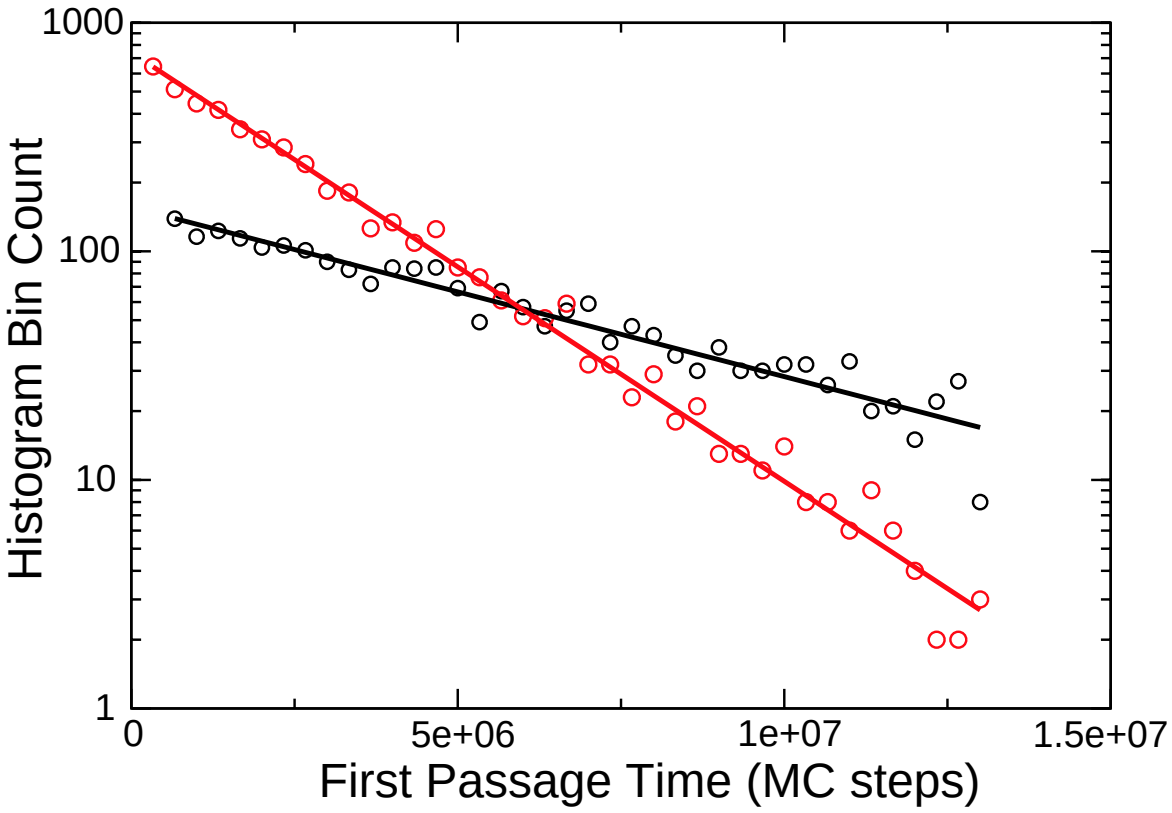

Figure 3.5: The distributions of first passage times for folding (black) and unfolding (red) observed during a 2.7×10^{10}-step kinetic MC at 475 K. Approximately 4700 folding and unfolding events were observed. A folding first passage time is defined as the time elapsed from when the particle enters the unfolded region from the buffer region (having previously been in the folded region), until it re-enters the folded region. The unfolding first passage time is defined similarly. The semi-log plot of the histograms of the first passage times is shown as circles, while the lines represent the best-fit exponential curve.

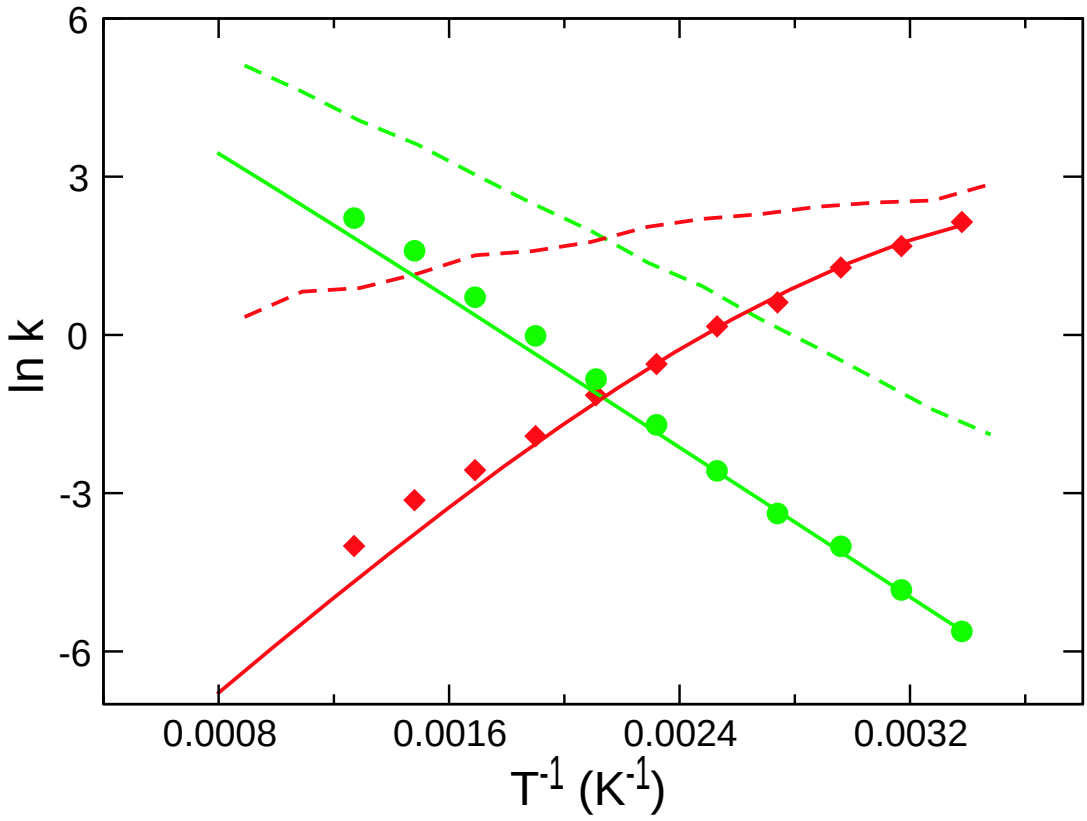

Figure 3.6: The temperature dependence of the folding and unfolding rate constants (solid lines and symbols). Folding and unfolding rates are indicated by red and green color, respectively. The folding and unfolding rates corresponding to a system with a smaller activation energy for folding (Appendix I) are shown for comparison (dashed lines). The rate constants plotted in symbols were derived from kinetic MC simulations run at different temperatures. The solid lines represent the rates calculated using the Arrhenius equation based on activation energies derived from the PMF along x (Figure 3.3). Rate constants are expressed in units of 10^{-6} per MC step.

choosing a move set for which equilibration along the y coordinate is faster than along the x coordinate. The alternative potential parametrization in Appendix I, which is characterized by a smaller average potential energy of the unfolded state relative to the folded and the transition states, leads to a weaker temperature dependence of the folding rate (Figure 3.6, dashed lines). Since the slope of the Arrhenius curve is proportional to the activation energy, this difference of the rates is consistent with the smaller energy of activation obtained with the alternative parametrization.

The folding rates decrease with increasing temperature, a phenomenon which has been observed in the kinetics of protein folding[34, 105, 36, 37, 38]. Processes displaying anti-Arrhenius behavior are said to be characterized by a negative effective activation energy, whereby the enthalpy of the unfolded state is larger than that of the transition state. The free energy of activation of these processes, however, remains positive due to the activation entropy favoring the unfolded state. The negative activation entropy is associated with the smaller number of accessible conformations at the transition state relative to the unfolded state; that is, the transition state constitutes an entropic "bottleneck" that needs to be traversed for the transition to the folded state to occur. These elements clearly exist in the simplified two-dimensional system under investigation. Since the potential energy decreases monotonically from the unfolded state to the folded state, the average potential energy at the transition state ($x = 0$) is smaller than the average potential energy of the unfolded state, leading to the observed anti-Arrhenius behavior of the rate of folding. Despite the enthalpic driving force favoring the transition state, the free energy of activation for folding remains positive at all temperatures examined (as the calculated PMF along the x coordinate shows). This is because the entropy of the transition state is smaller than the entropy of the unfolded state due to the larger accessible configuration space along the y coordinate (Figure 3.1). The entropic destabilization of the transition state, which (as in protein folding) can be described as acting as a "bottleneck", more than offsets the enthalpic stabilization, leading to the observed positive activation free energy for folding.

Often the observed folding rates of proteins show non-monotonic behavior with respect to the temperature; the folding rate increases with temperature at low temperatures as in normal Arrhenius behavior, switching to anti-Arrhenius behavior at higher temperatures, when the folding rate decreases with increasing temperature. This phenomenon is rationalized in terms of a negative activation heat capacity. The activation heat capacity is defined as the temperature derivative of the activation energy, and a negative value of the activation heat capacity indicates that the unfolded state has a larger heat capacity than the transition state. The observed negative heat capacity of activation of protein folding has been variously interpreted as being due to the hydrophobic effect[34, 105] or to the difference of the distribution of energies of the molecular conformations experienced as a function of temperature[40, 36]. The curvature of the Arrhenius plot is related to the activation heat capacity. The present simplified two-dimensional system does not have a large enough heat capacity of activation to reproduce this turnover from Arrhenius to anti-Arrhenius behavior within the temperature range we have investigated. Thus, the results extracted from this model are applicable only to the anti-Arrhenius temperature regime of the protein folding process.

Figure 3.7 shows the number of direct round trip transition events N_{direct} observed during MC simulations of $N_{MC} = 5 \times 10^9$ steps as a function of temperature. We use the number of transitions as a measure of the efficiency of conformational sampling, which determines the rate of convergence of thermodynamic quantities extracted from the simulations. The results of Figure 3.7 show that conformational sampling efficiency of the uncoupled simulation varies non-monotonically with the temperature. There is a 40-fold increase in transitions from 300 K to 474 K, the temperature at which the maximum is observed. This decreases for temperatures higher than 474 K, reaching a 10-fold reduction at 800 K (relative to the maximum). As the results in Figure 3.7 show, this behavior mirrors almost exactly the behavior of the harmonic mean $(k_f^{-1} + k_u^{-1})^{-1}$ of the folding and unfolding rates (from Figure 3.6) as a function of temperature (we note that our use of the

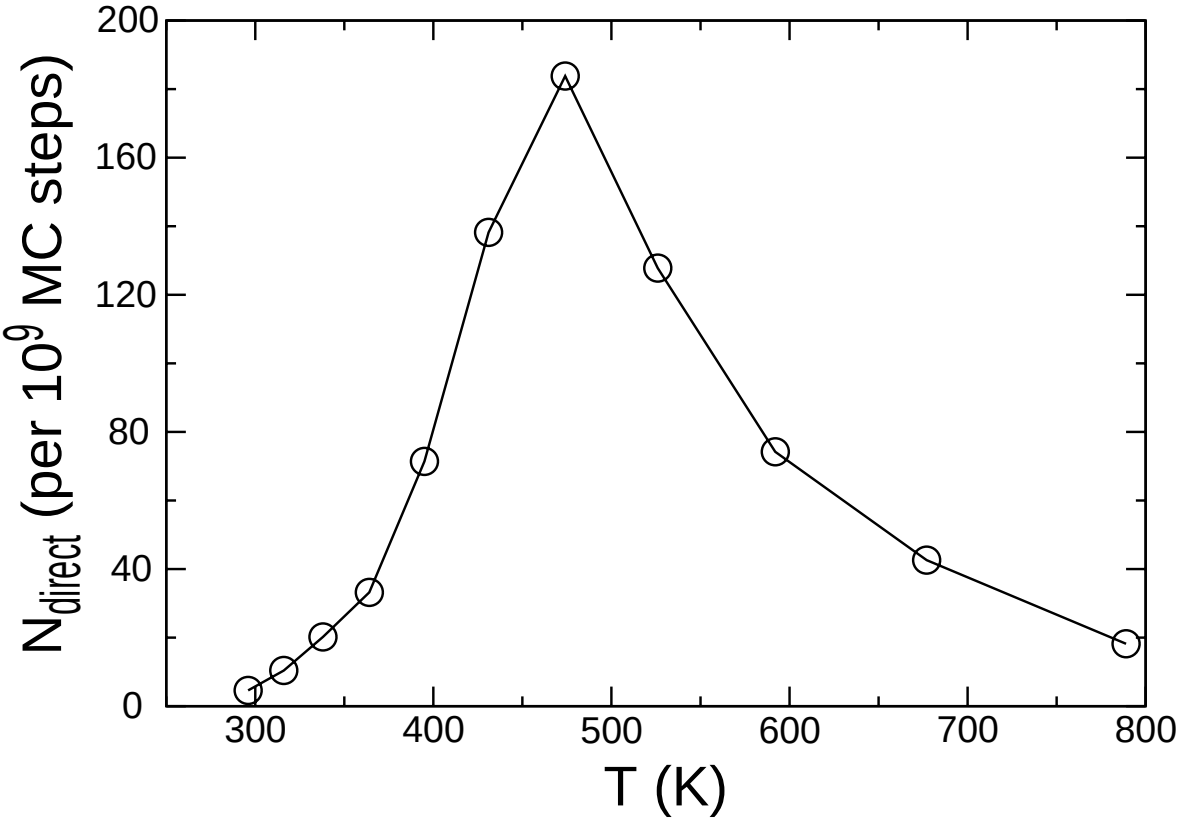

Figure 3.7: Number of direct round trip transition events N_{direct} in single temperature uncoupled simulations over the temperature range $296 - -789\ K$ in 5×10^9 MC steps. The curve plotted as a solid line was calculated from the harmonic mean of the folding and unfolding rates estimated from the mean of the folding and unfolding first passage time distribution (Figure 3.5) obtained by MC simulations at each temperature, while the number of events counted directly from the MC simulations at individual temperatures are plotted as circles. The high level of agreement indicates that the system is very well approximated as a two-state activated process.

term “harmonic mean” differs from standard usage by a factor of 2, which is natural given that we are considering a round-trip, i.e. a single “transition event” involves two conformational transitions). The agreement between the harmonic mean of the rates and the number of direct round trip transitions is expected for a two-state activated equilibrium, since the average time of a round-trip excursion from the folded to the unfolded state and back is the sum of the average folding and unfolding times $\tau_f = k_f^{-1}$ and $\tau_u = k_u^{-1}$, respectively: $N_{\mathrm{direct}} = N_{MC}/(\tau_f + \tau_u)$.

3.3.2 RE simulations using MC on the continuous potential

In chapter2, we analyzed the convergence and efficiency of replica exchange using a discrete model for folding and unfolding. We found that when the physical kinetics shows anti-Arrhenius temperature dependence, there exists an optimal maximal temperature beyond which the efficiency of the replica exchange method is degraded. Similar behavior is expected from RE simulations using the continuous two-dimensional potential, with possible differences arising from the more complex nature of the present model, where the folded and unfolded states have internal structure. We performed replica exchange simulations on the continuous two-dimensional potential with MC as the dynamic propagator, and replica exchange proposals made periodically between adjacent temperatures every N_X MC steps. The efficiency of conformational sampling was monitored by counting the number of temperature-conditional transition events N_{TE} defined in section 3.2.3 above.

In order to directly compare with the results obtained previously, we first performed replica exchange using two replicas. Although such a simulation would not be realistic in general for a protein system due to poor energy overlap and very inefficient temperature exchange, it is feasible in the two-dimensional potential. The result for a 2×10^9-step simulation where the lower temperature is held fixed at 296 K and the upper temperature varies from 296 to 789 K is shown in Figure 3.8 (green, red and blue dots). We see behavior

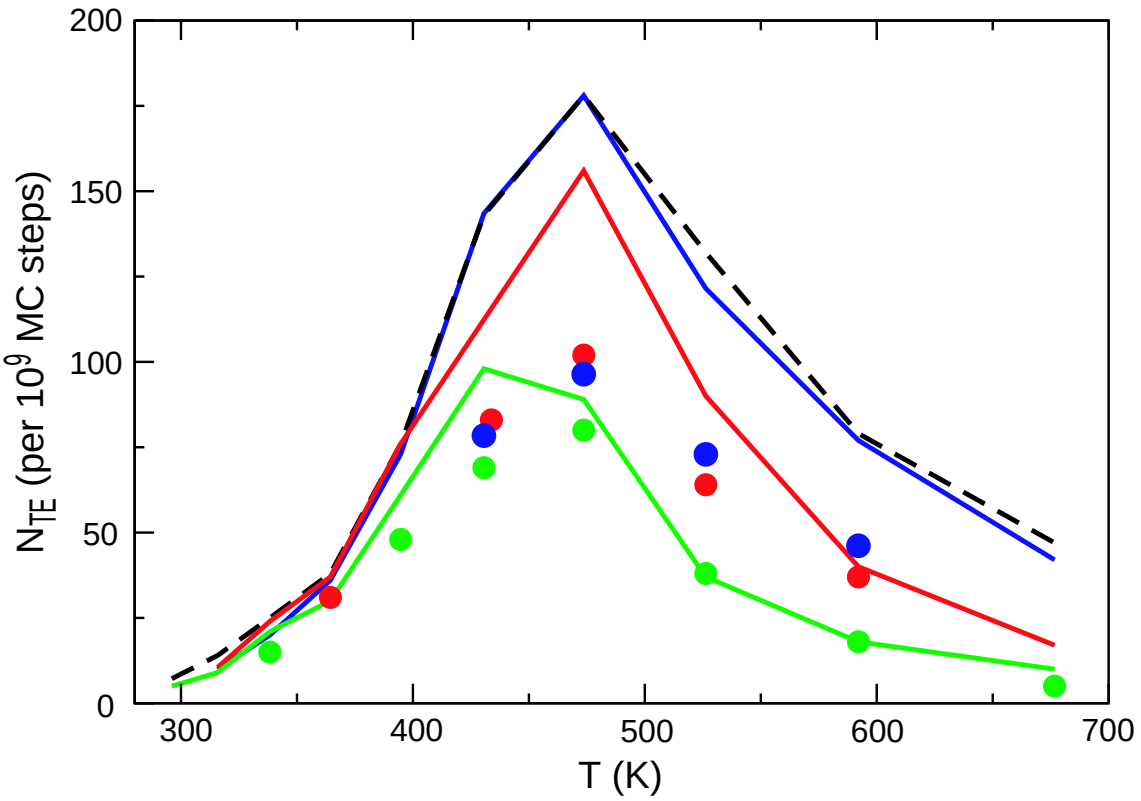

Figure 3.8: The dependence of the number of temperature-conditional transition events N_{TE} (section 3.2.3) on the temperature of the high-temperature replica for a two-replica simulation on the continuous potential (circles), and comparison with predicted transition events using discrete NRE model (Figure 3.2) (curves). For all simulations, one replica was fixed at 296 K and the other replica was scanned from 296 K to 789 K. The black dashed line corresponds to the discrete model prediction in the large-α limit. The solid curves are the predicted N_{TE} using the NRE model with temperature dependent folding and unfolding rates taken from the kinetic MC simulations on the continuous potential (shown in Figure 3.6) and using an α corresponding to an attempted temperature exchange. The circles are the N_{TE} values observed in 2×10^9 MC step RE simulations on the continuous potential. The green, red, and blue data correspond to N_X values of 1 000, 200, and 20, respectively.

similar to that seen for the discrete model studied previously: the number of temperature-conditional transitions N_{TE} has non-monotonic behavior and exhibits a maximum at an optimal high temperature given by the maximal harmonic mean of the folding and unfolding rates (474 K). This maximum point is approximately independent of the rate at which attempted temperature exchanges occur. While the location of the maximum is in agreement with our previous results[98], the magnitude of the number of transition events is not. We have shown that for NRE simulations employing a two-state model (folded and unfolded states), the number of transition events is given by the average over all temperatures of the harmonic means of the folding and unfolding rates, provided that the rate of temperature exchanges is sufficiently fast[98]. In the continuous model, we find that the number of transitions is significantly lower than that predicted from the average of the harmonic means of the rates (Figure 3.8, black dashed line). This may be due to the finite rate of temperature exchanges, deviations from the pure Markovian kinetics of the two-state discrete model, or a combination of these effects.

To test whether this reduced number of transitions is due to insufficiently fast temperature exchange attempts, we performed several simulations in which we varied N_X (the number of MC steps between attempted temperature exchanges). We see in Table 3.1 that N_{TE} is approximately constant provided that the attempted exchange rate is faster than a critical value of $N_X \approx 500$. For less frequent exchange attempts, we see a substantial decrease in the number of transitions. Thus, the number of unfolding and refolding transitions cannot be increased simply by increasing the rate of attempted exchanges.

3.3.3 Non-Markovian effects revealed by comparison of continuous and discrete RE simulations

To explore causes for the observed transition deficit, we performed simulations using the discrete NRE model (Figure 3.2) using kinetic parameters derived from the two-dimensional

continuous potential (Figure 3.6). To map the rates determined using the continuous potential to the discrete model, we used the folding and unfolding rates directly, expressed in units of 10^{-6} per MC step. Different values of α were used for the $F_1F_2 \rightleftharpoons F_2F_1$ and $U_1U_2 \rightleftharpoons U_2U_1$, and were set to $10^6/N_X$ multiplied by the empirical acceptance rate when both replicas are in the folded or unfolded state (0.853 and 0.395, respectively).

If we compare the observed number of transitions seen in the continuous model with the number predicted by the NRE model with the same rate parameters (Table 3.1) we see that there is good agreement when the attempted exchange rate is small, but substantial disagreement when it becomes larger. In particular, while the number of transitions using the continuous model reaches a plateau value at $N_X \approx 1\ 000$, the predicted number of transitions in the NRE model continues to increase, asymptotically approaching the value predicted by the average of harmonic means. Similarly, comparison of the predicted and observed number of transitions as a function of temperature (Figure 3.8) show a significant overestimation of the transition rate by the NRE model, and that this overestimation is much more severe when the rate of attempted temperature exchanges is fast. For example, while the N_{TE} predicted from the NRE model has essentially reached the asymptotic limit when $N_X = 20$ (blue curve), the observed N_{TE} values are essentially unchanged relative to those obtained when $N_X = 200$ (compare blue and red circles). The continuous two-dimensional model thus appears to contain an inherent "speed limit" which prevents it from achieving the transition rates expected for a fully Markovian system, even if the temperature exchanges are attempted frequently.

One possible origin of this speed limit is that the average effective rates are different in the coupled and uncoupled systems. To test this, we analyzed the kinetics of the continuous RE simulation by using the NRE model to "reverse-engineer" the apparent rates by estimating the mean residence times and branching ratios for various RE macrostates. If the system is Markovian, then the rate k_{tot} given by the inverse of the mean residence time is the sum of the rates exiting that state. The rate corresponding to a given edge can

Table 3.1: Number of temperature-conditional transition events in 2×10^9 MC steps for two replicas (with temperatures of 296 K and 474 K) as a function of the number of MC steps between attempted temperature exchanges (N_X), and observed temperature-conditional mean first passage times (in units of 10^6 MC steps, see text for details).

N_X	N_{TE} per replica		Temperature-conditional MFPTs	
	observed (continuous)	predicted (NRE)	$F_1 \rightarrow U_1$	$U_1 \rightarrow F_1$
10 000	22	24	91.8	5.6
2 000	52	73	31.6	5.7
1 000	80	105	23.1	5.7
500	93	134	19.1	5.8
200	102	162	16.0	5.7
100	99	168	14.3	5.9
80	98	172	14.6	5.8
50	98	176	14.8	5.7
20	96	177	14.9	6.1
0[a]	—	178	—	—

[a] Predicted N_{TE} based on the harmonic mean relationship for the $\alpha \rightarrow \infty$ limit.

Table 3.2: Empirical "reverse-engineered" rates at temperatures $T_1 = 296$ K and $T_2 = 474$ K (in units of 10^{-6} MC step) from continuous potential simulation data assuming the network topology of Figure 3.2 .

	uncoupled	reverse-engineered rates			
	rates	$N_X = 10\,000$	$N_x = 2\,000$	$N_X = 200$	$N_X = 100$
k_{f1}	6.08	5.66	6.10	5.27	6.33
k_{u1}	0.0036	0.0038	0.0036	0.0037	0.0037
k_{f2}	0.279	0.288	0.299	0.290	0.306
k_{u2}	0.420	0.420	0.419	0.427	0.425

then be estimated by multiplying k_{tot} by the fraction of residences that exit via that edge (the branching ratio). The results are shown in Table 3.2. The reverse-engineered rates generally agree with the uncoupled folding and unfolding rates estimated from kinetic MC, and this is true both for rapid and slow attempted temperature exchange rates. Therefore, the temperature exchanges do not perturb the average kinetics of the system, and cannot be a cause of the limit on the transition rates at rapid temperature exchange rates.

In order to further investigate the origin of the observed speed limit, we calculated the mean first passage times (MFPTs) for temperature conditional folding and unfolding, i.e. the average time for a replica unfolded at low temperature to become folded at low temperature (regardless of path), or *vice versa*. The resulting MFPTs for the continuous potential are shown in Table 3.1. We see there that the N_{TE} speed limit arises exclusively from a limitation in the fastest achievable unfolding rate, since the folding process is independent of N_X and is not rate limiting. This can be understood by noting that the values of α corresponding to the N_X values used are at least two orders of magnitude larger than the folding and unfolding rates. To unfold, the system need only make use of temperature exchange transitions that correspond to α (i.e. the solid cyan arrows of Figure 3.2). Since α is already much larger than the other rates, changes to it due to changes in N_X will not significantly change the MFPT for folding.

On the other hand, the unfolding process (if it occurs via an indirect route, which is likely given the very small value of k_{u1}) requires the system to use a "$w\alpha$ edge" (i.e. a dashed cyan arrow in Figure 3.2). Since $w \approx 10^{-4}$ for the temperatures used here, $w\alpha$ is now slower than or comparable to the folding and unfolding rates, and therefore changes in N_X can make a substantial impact on the unfolding MFPT. Thus, the N_{TE} speed limit can be traced to the kinetics of temperature conditional unfolding, and must arise from some difference between the unfolding kinetics in the continuous potential and the fully Markovian NRE model.

One obvious way in which the continuous and NRE models differ is that the macrostates

in the continuous potential have spatial extent, unlike the NRE states which lack internal structure. This means that a finite time is required for the particle to transit the non-equivalent microstates that make up the two wells. In fact, we observe that the correlation time for diffusion in the x direction *in the unfolded well* at 474 K is approximately 1 400 MC steps. This timescale is of the same magnitude as the N_X value at which the speed limit effect of Table 3.1 begins to occur, suggesting that there may in fact be a connection between the observed N_{TE} speed limit and conformational diffusion within the free energy wells. Such dependence of the kinetics on the internal structure of the macrostate can lead to non-Markovian behavior.

Formally, a process is Markovian if and only if the observed propagators (Green's functions) do not depend on the history of the trajectory prior to the current state, i.e.

$$P(x_3, t_3|x_1, t_1; x_2, t_2) = P(x_3, t_3|x_2, t_2) \tag{3.4}$$

for all states x_1, x_2, x_3 and all times $t_1 < t_2 < t_3$. Although equation 3.4 could be used to directly detect deviations from Markovian behavior, previous work has typically used other analysis methods to detect such deviations[79, 112, 113]. For example, in a Markovian process the rate matrix $\mathbf{K}$ determines the propagators via the master equation

$$\dot{\mathbf{p}}(t) = \mathbf{K}\mathbf{p}(t), \tag{3.5}$$

where $\mathbf{p}(t)$ is the vector of propagators at time t. The formal solution of Equation 3.5 is given by $\mathbf{p}(t) = e^{\mathbf{K}t}\mathbf{p}(0)$, and therefore $e^{\mathbf{K}\tau}$ can be thought of as a transition matrix $\mathbf{T}(\tau)$, i.e. the matrix of probabilities of being in state x_j at time τ given that the system was in state x_i at time 0. If we denote the eigenvalues of $\mathbf{K}$ by $\lambda_1 > \lambda_2 > \cdots$ and the eigenvalues of $\mathbf{T}(\tau)$ by $\mu_1(\tau) > \mu_2(\tau) > \cdots$, then $\mu_i(\tau) = e^{\lambda_i \tau}$. This can be used as a test of Markovian behavior, since $\mathbf{T}(\tau)$ can be empirically estimated from a trajectory. Different values of the lag time τ will yield different values of $\mu_i(\tau)$, however $\tau / \ln \mu_i(\tau)$ should be independent of τ if the kinetics is Markovian[79, 113]. Alternatively, the Markov property

can be tested by analyzing the transition probabilities as a function of lag time using an information theoretic measure based on Shannon's entropy[112].

We have chosen to detect deviations from Markovian kinetics by examining the observed residence time distributions and branching ratios, which provides insights into the physical origin and the mechanism by which the non-Markovian effects enter into the stochastic process. In our simulations on the continuous potential, we have found that the residence time distributions in the macrostates are exponential to within statistical uncertainty (data not shown), and thus by themselves are consistent with Markovian kinetics. The branching probabilities, however, are significantly dependent on the preceding macrostate. We focused on transitions entering and leaving the thermodynamically favored U_2F_1 macrostate (or its symmetry-related state F_1U_2). We ran a several trajectories using different rates of attempted temperature exchange and tallied the number of times each macrostate sequence (X, U_2F_1, Y) was observed in each (where $X, Y \in \{F_2F_1, U_2U_1, U_1F_2\}$). These counts were transformed into normalized branching probabilities, where $P(X|Y)$ denotes the history-independent branching probability of next visiting macrostate X given that the system is currently in macrostate Y, and $P(X|Z, Y)$ denotes the history-dependent branching probability of next visiting macrostate X given that the system is currently in macrostate Y and had been in macrostate Z immediately prior (Table 3.3).

If the kinetics is Markovian, then the history-dependent and the corresponding history-independent branching probabilities will be equal:

$$P(X|Z, Y) = P(X|Y),$$

from which it follows that history-dependent branching probabilities that differ only in the history condition will also be equal:

$$P(X|Z, Y) = P(X|W, Y).$$

This is clearly not the case for the data in Table 3.3. For example, we see that the history-dependent branching probabilities $P(U_1F_2|F_2F_1, U_2F_1)$, and $P(F_2F_1|F_2F_1, U_2F_1)$ differ

significantly from their corresponding history-independent branching probabilities $P(U_1F_2|U_2F_1)$ and $P(F_2F_1|U_2F_1)$, and the branching probability $P(U_1F_2|F_2F_1, U_2F_1)$ is significantly smaller than $P(U_1F_2|U_1F_2, U_2F_1)$. This is most pronounced when the rate of attempted temperature exchanges is fast.

Examination of the kinetic scheme of Figure 3.2 indicates that the deviations from Markovian behavior seen in Table 3.3 are consistent with a reduction in the number of temperature-conditional round-trip conformational transition events. If the unfolding rate at low temperature is negligible, then a low-temperature folded conformation unfolds predominantly via indirect paths of the form $F_1F_2 \rightarrow F_2F_1 \rightarrow U_2F_1 \rightarrow U_1F_2$ or $F_1U_2 \rightarrow F_2U_1 \rightarrow U_2U_1 \rightarrow U_1U_2$. In the former case, the $F_2F_1 \rightarrow U_2F_1$ step is more likely to be reversed when the temperature exchange rate is rapid (Table 3.3), as is the $F_1U_2 \rightarrow F_2U_1$ step in the latter case (which follows by symmetry from the $U_2F_1 \rightarrow U_1F_2$ results of Table 3.3). Thus, increasing the rate of attempted temperature exchanges increases the probability of counterproductive backtracking relative to the Markovian case, resulting in a decrease in the rate of temperature-conditional unfolding events, and therefore at corresponding decrease in N_{TE} (since temperature-conditional unfolding was shown above to be rate-limiting).

Although the results presented here do not identify the physical origin of the non-Markovian kinetics, we hypothesize that it is due to the finite time required for diffusion of the particle within the macrostates. This effect does not arise in the NRE model, since in there the macrostates have no internal structure, and the probability of making a transition to a given macrostate at any instant dt is the same, regardless of which macrostate the system was in previously or how long it has been in the current macrostate. The behavior of the continuous system within the wells is not Markovian, since the system has memory that is mediated by conformational diffusion within the macrostate. This correlation in time of the particle's position (and energy) implies that there is a maximal effective value of the rate of statistically independent temperature exchanges, which is limited by the time

required for conformational relaxation *within* the folded and unfolded macrostates.

3.3.4 Dependence of RE efficiency on the number of replicas

The above results were obtained with two replicas, which is not typical for replica exchange simulations that would be carried out for peptides and proteins. To investigate the effect of adding additional replicas, we performed a series of simulations of 2×10^9 MC steps with 2 to 15 replicas distributed uniformly in T^{-1} from 296 to 789 K. The results are shown in Figure 3.9. One important issue that arises when considering such a set of results is the appropriate measure of conformational sampling efficiency of RE. If we consider the total number of transition events N_{TE} (direct and indirect) in all replicas, then we would see for the most part a monotonic increase of efficiency as a function of the number of replicas N simply because the number of indirect "channels" for transitions is linearly increasing. This measure of efficiency, however, implicitly assumes that computer power is inexpensive and that the convergence rate of the simulation is the important limiting factor. If both computer resources and the convergence rate are limiting factors, a more appropriate measure is the computational efficiency calculated as the number of transition events per replica (N_{TE}/N). According to this measure, a replica exchange simulation with $N + 1$ replicas is considered more efficient than one with N replicas only if the introduction of the additional replica provides more than a proportional increase in the number of transition events at the temperature of interest.

We find that the efficiency increases strongly as a function of N when N is small, reaches a maximum, and decreases with N for larger N (Figure 3.9). This pattern is unchanged as a function of the rate of attempted temperature exchanges, showing a scaling approximately consistent with the results in Table 3.1. The trends seen here are qualitatively similar to that seen previously in the NRE two-state discrete model[98] with finite α. In that work, we attributed the decrease with increasing number of replicas beyond an

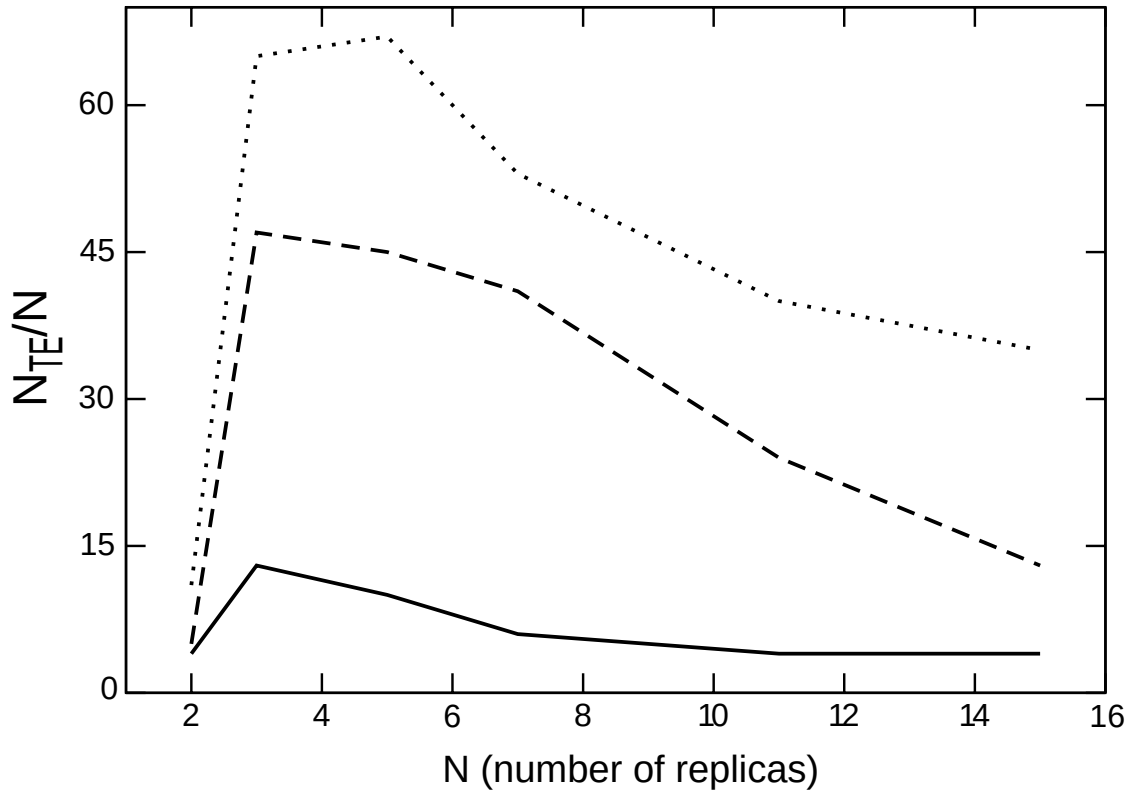

Figure 3.9: Number of transition events N_{TE} (section 3.2.3) normalized by the number of replicas in 2×10^9 MC steps for 2 to 15 replicas exponentially distributed in temperature from 296 to 789 K. Temperature exchanges were attempted every 10 000 (solid), 1 000 (dashed), and 200 (dotted) MC steps.

optimum value in part to a combinatoric effect that decreases the relative size of the "target" space of configurations in which a replica is at the temperature of interest relative to the total temperature/configuration space. It is reasonable to assume that a similar effect is occurring here as well. We will address this in a future communication.

The results in Figure 3.9 were obtained with a relatively uniform distribution of temperatures. It is of interest to consider the effect on efficiency of changing that temperature distribution. In our previous work[98], we concluded that in the context of the discrete network model in the "large α" limit, the optimal temperature distribution is one replica at the temperature of interest, and the rest at the temperature which maximizes the harmonic mean of the folding and unfolding rates. That model, however, was limited in its realism in that it did not have explicit energy distribution functions. Furthermore, it is clear from the results presented in the previous section that very large effective values of α may not be achievable in real systems. The continuous two-dimensional potential studied here provides a better test system for studying these questions.

In Figure 3.10 we show the relative number of temperature-conditional transition events in 2×10^9 MC steps for three different temperature distributions of 11 replicas: (A) uniformly distributed in T^{-1} from 296 to 789 K, (B) 6 replicas uniformly distributed in T^{-1} from 296 to 474 K (the optimal temperature) and the remaining 5 "bunched up" at the optimal temperature, and (C) 5 replicas bunched up at the optimal temperature with the remaining distributed in the 296 to 474 K range but strongly skewed toward the optimal temperature. Temperature distribution B provides more than a 50% increase in efficiency relative to the uniform distribution over the large temperature range. This is consistent with our discrete model results, and indicates that it is possible to include temperatures that are "too high" when the system exhibits anti-Arrhenius kinetics. However, we can increase the efficiency even further (to more than a factor of 2.5 over the baseline result) by skewing the temperature distribution to increase the number of replicas in the vicinity of the transition temperature (distribution C). Previous work by Hansmann et al. has suggested that such

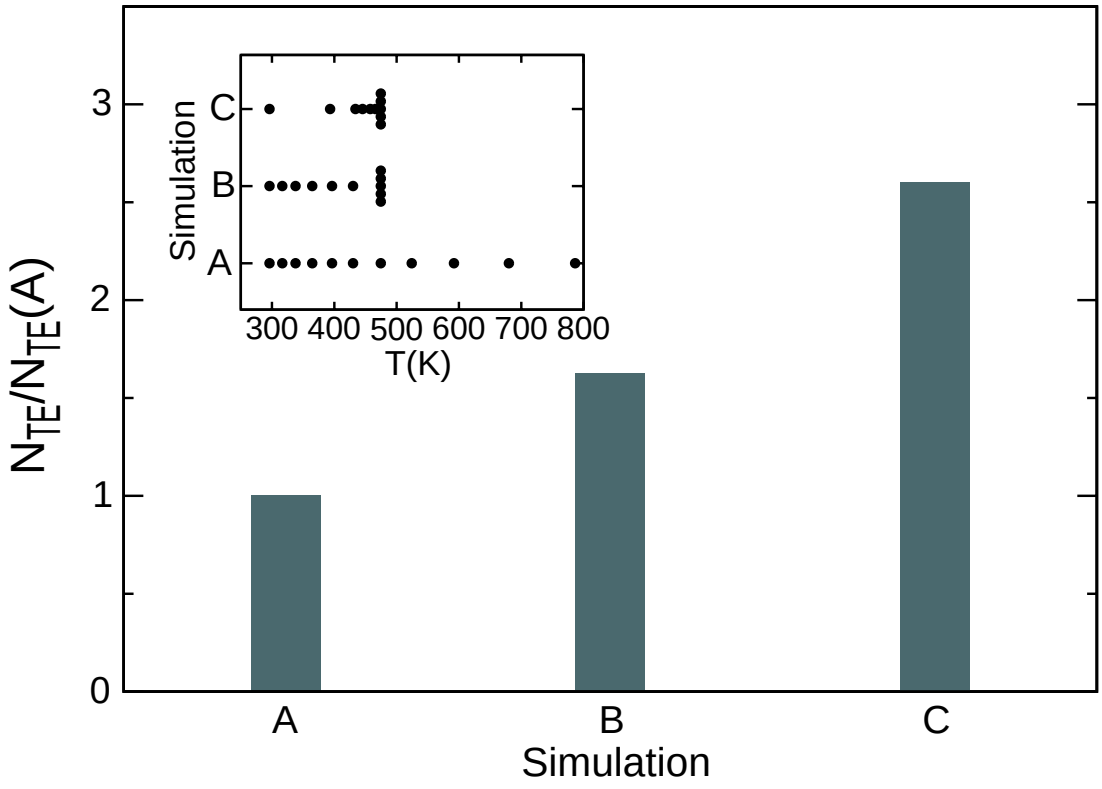

Figure 3.10: Number of transition events N_{TE} (section 3.2.3) observed in 2×10^9 MC steps for three different 11-replica RE simulations performed using the continuous potential with $N_X = 200$. The temperature distributions for the three simulations are shown in the inset. Transition counts have been normalized by the N_{TE} of simulation A.

concentration of the temperatures near a bottleneck can improve temperature mixing[29]. However, the improved efficiency may simply be due to the increased number of replicas near the optimal temperature. The clarification of the relative contributions from these two effects will also be addressed in a future communication.

3.4 Conclusions

One of the challenges of studying the computational efficiency of replica exchange has been the difficulty in running molecular simulations sufficiently long to obtain full convergence and meaningful statistics. This is particularly daunting if such simulations must be run multiple times to assess the effect of differences in simulation protocols and parameters. The use of simplified model systems allows for thorough theoretical, conceptual, and computational analysis of the problem that can provide insights into the factors that limit the efficiency of replica exchange in more realistic molecular systems.

Our previous work made use of a highly simplified discrete model for protein folding with two conformational states at several temperatures[98]. While this system did provide useful insights, it was limited in a number of ways, and in particular was fully Markovian. Here we have described a two-dimensional continuous potential function and an associated move set that allows us to perform MC and replica exchange MC simulations in a system that is small enough to quickly converge but yet is rich in complexity that is reminiscent of molecular systems. While many of the results are consistent with those observed previously, novel effects are also seen. In particular, we have confirmed that the efficiency of replica exchange in more complex systems is fundamentally limited by the timescale of conformational diffusion within basins, as we had anticipated[98]. We expect that such behavior will also be present (perhaps even more strongly) in molecular systems.

There are many unresolved questions raised by this work. One question for which our two-dimensional system would be a good model is for studying the relationship between

conformational and thermal diffusion. Optimization of the diffusion of replicas in temperature space has been a major focus of recent theoretical and computational study of the replica exchange method[12, 24, 25, 26, 27, 28, 29, 99]. However, the convergence of thermodynamic quantities is not limited by thermal diffusion *per se*, but by the exploration of the conformational space of the system. While very poor thermal diffusion obviously defeats the purpose of replica exchange by effectively reducing it to a set of parallel uncoupled simulations, it is not clear that further optimization of thermal diffusion that is already "reasonably good" will automatically improve convergence. Some recent work has begun to address the role of basin-to-basin transitions[30, 31]. Similarly, some work on the optimization of thermal diffusion has emphasized the role of temperature bottlenecks[29], which may turn out to be fundamentally conformational in nature. The exact relationship between thermal and conformational diffusion remains to be fully clarified, and we look forward to studying this and other questions using simplified continuous and discrete models of replica exchange.

3.5 Appendix I —-The alternative potential function

The alternative potential with decreased average potential energy differences between folded, unfolded and transition states is of the same general form as the primary potential described in the Methods section and Figure 3.1, but with the boundary function parameters $\delta = 10^{-5}$, $b = 1$, and $n_1 = 3.5$ and potential energy

$$U(x,y) = \begin{cases} a_1(x+x_0)^2 + b_1 y^2, & -1 \le x < -x_1 \quad 0 \le y \le B(x) \\ -a_2 x^2 + b_1 y^2 + c_0, & -x_1 \le x \le 0, \quad 0 \le y \le B(x) \\ a_3 x^2 + b_1 y^2 + c_0, & 0 < x \le 1, \quad 0 \le y \le B(x) \\ \infty & \text{otherwise} \end{cases},$$

with $a_1 = 25$ kcal/mol, $a_2 = 250$ kcal/mol, $a_3 = 10$ kcal/mol, $b_1 = 1000$ kcal/mol, $c_0 = 6$ kcal/mol. The constants x_0 and x_1 were the same as for the primary potential. curves of

these two potentials, the alternative one is much less steep than the other.

3.6 Appendix II: Publication attached

Part of the contents of this chapter was published in *J. Phys. Chem. B*, 112, 6083-6093(2008).

Table 3.3: History dependent and independent branching probabilities from state U_2F_1.

conditional probability	maximum likelihood estimate[a] (and 95% credible interval[b]) $N_X = 200$	$N_X = 10\,000$
$P(U_1F_2\|U_1F_2, U_2F_1)$	0.906 (0.904, 0.908)	0.168 (0.144, 0.195)
$P(U_1F_2\|F_1F_2, U_2F_1)$	0.521 (0.153, 0.530)	0.094 (0.088, 0.101)
$P(U_1F_2\|U_2F_1)$	0.849 (0.846, 0.851)	0.103 (0.096, 0.110)
$P(F_2F_1\|F_2F_1, U_2F_1)$	0.477 (0.469, 0.486)	0.895 (0.888, 0.902)
$P(F_2F_1\|U_1F_2, U_2F_1)$	0.092 (0.090, 0.094)	0.816 (0.788, 0.841)
$P(F_2F_1\|U_2F_1)$	0.150 (0.147, 0.152)	0.886 (0.878, 0.893)

[a] Maximum likelihood estimates determined using $P(a_1|b,c) = \#(b,c,a_1)/\sum_i \#(b,c,a_i)$ and $P(a_1|b) = \sum_i \#(c_i,b,a_1)/\sum_{jk} \#(c_j,b,a_k)$, where $\#(i,j,k)$ is the number of occurrences of the ordered triple (i,j,k).

[b] Bayesian credible intervals under a uniform prior given by the 0.025 and 0.975 quantiles of the distribution $P(p) \propto p^n(1-p)^{N-n}$, where n and N are the numerator and denominator, respectively, of the fraction used to calculate the maximum likelihood estimate.

Simple Continuous and Discrete Models for Simulating Replica Exchange Simulations of Protein Folding†

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: August 8, 2007; In Final Form: November 8, 2007

The efficiency of temperature replica exchange (RE) simulations hinge on their ability to enhance conformational sampling at physiological temperatures by taking advantage of more rapid conformational interconversions at higher temperatures. While temperature RE is a parallel simulation technique that is relatively straightforward to implement, kinetics in the RE ensemble is complicated, and there is much to learn about how best to employ RE simulations in computational biophysics. Protein folding rates often slow down above a certain temperature due to entropic bottlenecks. This "anti-Arrhenius" behavior represents a challenge for RE. However, it is far from straightforward to systematically explore the impact of this on RE by brute force molecular simulations, since RE simulations of protein folding are very difficult to converge. To understand some of the basic mechanisms that determine the efficiency of RE, it is useful to study simplified low dimensionality systems that share some of the key characteristics of molecular systems. Results are presented concerning the efficiency of temperature RE on a continuous two-dimensional potential that contains an entropic bottleneck. Optimal efficiency was obtained when the temperatures of the replicas did not exceed the temperature at which the harmonic mean of the folding and unfolding rates is maximized. This confirms a result we previously obtained using a discrete network model of RE. Comparison of the efficiencies obtained using the continuous and discrete models makes it possible to identify non-Markovian effects, which slow down equilibration of the RE ensemble on the more complex continuous potential. In particular, the rate of temperature diffusion and also the efficiency of RE is limited by the time scale of conformational rearrangements within free energy basins.

1. Introduction

One of the key challenges in the computer simulation of proteins at the atomic level is the sampling of conformational space. The efficiency of many common sampling protocols, such as Monte Carlo (MC) and molecular dynamics (MD), is limited by the lack of apparent ergodicity caused by high free energy barriers between conformational states and rugged energy landscapes. Replica exchange (RE) methods[1−5] are widely employed to enhance the conformational sampling efficiency of biomolecular simulations for the study of protein biophysics, including peptide and protein folding[6,7] and aggregation,[8−10] and protein−ligand interactions.[11,12] To accomplish barrier crossings, RE methods simulate a series of replicas over a range of potential parameters[13−17] or temperatures.[5] In the latter, replicas exchange temperatures following a Metropolis criterion designed to preserve canonical distributions. This scheme allows conformations at physiological temperatures, where conformational interconversions are rare, to switch to higher temperatures where transitions to other conformations are more likely. In a sense, therefore, the enhancement of conformational sampling at low temperatures is achieved by "borrowing'' the faster kinetics at higher temperatures.

The popularity of RE methods is due to their ease of implementation and their ability to enhance conformational sampling while preserving canonical distributions at the thermodynamic conditions of each replica. The properties of the RE algorithm and how it can be utilized most effectively for the study of protein folding and binding has received attention recently.[18−20] The determination of the temperature assignment and number of replicas to achieve optimal temperature mixing has been the subject of a variety of studies.[3,21−27] Recent work has also recognized the importance of conformational relaxation as a key limiting factor that can affect the efficiency of the RE algorithm.[18,19,26,28] While temperature RE is relatively straightforward to implement, kinetics in the RE ensemble is complicated and does not correspond in any simple way to the molecular kinetics (necessitating additional methods for the reconstruction of molecular kinetics from RE samples[29−32]). Molecular kinetics, however, can have a strong effect on RE, especially when the kinetics has complex temperature dependence. The anti-Arrhenius behavior typical of protein folding kinetics, where the folding rate above a critical threshold temperature decreases with increasing temperature,[33−36] is understood to occur when the transition state is energetically favored but entropically disfavored with respect to the reactants. Anti-Arrhenius behavior represents a challenge for temperature

† Part of the "Attila Szabo Festschrift".

* Corresponding author. 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/jp076377+ CCC: $40.75
Published on Web 02/06/2008

RE because, when folding exhibits anti-Arrhenius behavior, there exists a temperature (generally unknown) at which the folding and unfolding rates are optimal. If even higher temperatures beyond the optimal are included in the RE ensemble, this may degrade performance.[20]

Although some comparative studies aimed at determining the benefits of RE over conventional MD for peptide folding have been conducted,[19,37,38] it is far from straightforward to systematically explore the convergence properties of RE by brute force molecular simulations, since RE simulations of protein folding are very difficult to converge. To understand some of the basic mechanisms that determine the efficiency of RE, it is useful to study simplified low dimensionality systems that share some of the key characteristics of molecular systems. We recently investigated a discrete two-state network model for replica exchange (NRE), containing two conformational states (folded and unfolded) at each of several temperatures.[20] We found that the efficiency of RE for this system varies non-monotonically with respect to the temperature distribution of the replicas when the folding rate displays anti-Arrhenius behavior. The model showed that the rate of folding/unfolding events in RE is maximal when high-temperature replicas are placed near the temperature at which the harmonic mean of the folding and unfolding rates for the uncoupled system (k_f and k_u) is maximal. This result suggested that, in molecular simulations, adding high-temperature replicas does not necessarily lead to increased efficiency of exploration of conformational space, and that, instead, optimal efficiency could be obtained by placing replicas at specific temperatures determined by the temperature dependence of key kinetic rates of the system.

In this paper we extend this analysis by studying a continuous two-dimensional system designed to reproduce the anti-Arrhenius kinetics of a conformational equilibrium, such as a protein folding equilibrium, mediated by an entropic bottleneck. The two-dimensional system studied here is an extension of the potential model we originally used to study the convergence of the weighted histogram analysis method,[39] and is very similar in spirit to to the funnel-like golf course model for protein folding studied by Szabo and co-workers.[40] This two-dimensional system is sufficiently simple to be amenable to accurate analytical and numerical solution, while including some characteristics of molecular systems that were absent from the discrete NRE model. The present model is self-contained in that the kinetic rates are determined by the potential and the move set rather than being imposed, as in the NRE model of reference 20. Furthermore, and most importantly, the unfolded and folded macrostates have, like real molecular systems, microscopic internal structure. The new model makes it possible to follow the joint microscopic evolution of the system in conformational and temperature space. It incorporates the same discrete temperature exchange scheme commonly adopted in RE molecular simulations, and it allows us to study the effects of non-Markovian processes likely present in RE simulations of molecular systems.

In the next section we present the potential model and the kinetic scheme we have employed. We review the RE method and the NRE model we previously developed. We then summarize the thermodynamic and kinetic properties of the two-dimensional system and present results showing how these determine the efficiency of the RE method. The paper is then concluded by discussing the implications of these findings for RE simulations of molecular systems.

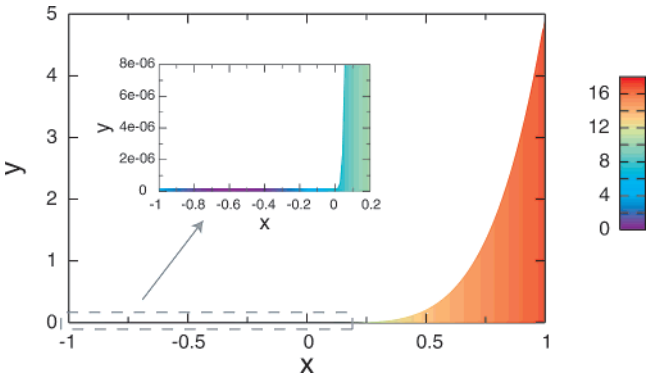

Figure 1. A schematic representation of the two-dimensional potential function used in this work. The colored area corresponds to the accessible region of the (x,y) plane, with the colors representing the magnitude of the potential energy at that (x,y) point (scale bar in kcal/mol). The potential energy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

2. Methods

2.1. The Two-Dimensional Continuous Potential. A two-dimensional potential was constructed to mimic the anti-Arrhenius temperature dependence of the folding rate seen in proteins. We designed this potential to have an energetic barrier when going from the "folded" to the "unfolded" 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 $-1 \le x \le 1$, $0 \le y \le B(x)$, where the boundary function $B(x)$ is a small constant for $x \le 0$ and an increasing function of x for $x > 0$ (Figure 1):

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

The use of a boundary of this form is based on a two-dimensional potential first used in our laboratory to study the convergence of the weighted histogram analysis method,[39] and is very similar in spirit to simplified models for protein folding studied by Bicout and Szabo[40] and the model of an entropic barrier by Zhou and Zwanzig.[41] The specific parameters δ, b, and n_1 were chosen together with the parameters of the potential function discussed below by trial and error to achieve a sufficiently strong temperature dependence to illustrate some of the possible consequences of anti-Arrhenius behavior on RE simulations. It is natural to choose the x axis to be the reaction coordinate, with $-1 \le x \le 0$ corresponding to the folded macrostate and $0 < x \le 1$ corresponding to the unfolded macrostate. The move set was chosen to be compatible with this reaction coordinate (see below). In order for folding and unfolding to be activated processes, however, it is necessary to add a potential energy function that has an energetic well as a function of x in the folded region, and increases with x in the unfolded region. Specifically, we use the potential function

$$U(x,y) = \begin{cases} a_1(x + x_0)^2 & -1 \le x < -x_1 \quad 0 \le y \le B(x) \\ -a_2x^2 + c_0 & -x_1 \le x \le 0 \quad 0 \le y \le B(x) \\ a_3x^{n_2} + c_0 & 0 < x < x_2 \quad 0 \le y \le B(x) \\ a_4x^{n_3} + c_1 & x_2 \le x \le 1 \quad 0 \le y \le B(x) \\ \infty & \text{otherwise} \end{cases} \tag{2}$$

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, $c_0 = 7.059$ kcal/mol, $b = 5$, $n_1 = 4.55$, $n_2 = 2$, $n_3 = 0.5$, and $\delta = 2 \times 10^{-7}$. The constants $x_0 = \sqrt{c_0(a_1+a_2)/a_1a_2}$, $x_1 = a_1x_0/(a_1 + a_2)$, $x_2 = (a_4n_3/a_3n_2)^{1/(n_2-n_3)}$, and $c_1 = c_0 - (a_4x_2^{n_3} - a_3x_2^{n_2})$ were chosen so that the first derivative of $U(x,y)$ is continuous. A graphical representation of the two-dimensional system studied here is shown in Figure 1.

2.2. Kinetics on the Two-Dimensional Continuous Potential. We use Metropolis MC sampling to simulate the movement of a particle in this two-dimensional potential. Kinetic MC has a long history in the study of protein folding using simplified models.[42−44] To ensure rapid equilibration along the y coordinate consistent with the choice of x as the reaction coordinate and because of the large size difference of the accessible region in the y direction between the folded and unfolded regions, we adopted an asymmetric MC proposal scheme,[39,45] in which the step size in the y direction is proportional to $B(x)$, i.e., a proposed move (x',y') is generated uniformly in the region $x - \Delta < x' < x + \Delta$, $y - B(x)\Delta < y' < y + B(x)\Delta$. The displacement parameter Δ was chosen such that the barrier crossing is slow but not prohibitively expensive and follows a linear regime (i.e., doubling Δ causes an approximate doubling in the number of barrier crossings). To correct for the asymmetric MC proposal distribution, the factor $\theta(|y' - y|/B(x')\Delta)$ was included to satisfy detailed balance, where $\theta(z)$ equals 1 if $z < 1$ and 0 otherwise.

Rate constants in units of MC steps were obtained via MC simulation by calculating the mean first passage times (MFPTs) between the two macrostates. The same displacement parameter $\Delta = 0.05$ was used for all temperatures. A "buffer region'' $-0.1 < x < 0.0437$ was defined as not belonging to either the folded or unfolded state to reduce artifactual rapid recrossings of the barrier.[46,47] For comparison, the temperature dependence of the folding and unfolding rate constants were also estimated from the potential of mean force (PMF) using the Arrhenius equation $k = A \exp(-\Delta G^{\dagger}/k_BT)$, where $\Delta G^{\dagger}$ is the free energy difference between the transition state and the appropriate macrostate. Free energies were extracted from the PMF along the x axis by averaging the PMF over the macrostates and transition region using numerical integration.

2.3. RE Simulation on the Two-Dimensional Continuous Potential. RE simulations were performed by running N MC simulations at N inverse temperatures $\beta_i = (k_BT_i)^{-1}$ ($\beta_1 > \beta_2 > ... > \beta_N$) in parallel. The state of the extended ensemble is specified by a joint configuration of N replicas $X = \{q_1,q_2,...,q_N\}$, where q_i is the configuration of replica i. Exchanges of configurations were attempted every N_X MC steps between pairs of replicas adjacent in temperature, and the attempted exchange $X = \{...,q_i,q_j,...\} \rightarrow X' = \{...,q_j,q_i,...\}$ was accepted with probability $w(X \rightarrow X')$. Given the potential energy function $U(q)$, the transition probability that satisfies detailed balance and reproduces the canonical ensemble is given by $w(X \rightarrow X') = \min\{1, \exp[-(\beta_j - \beta_i)(U(q_i) - U(q_j))]\}$.[5]

The efficiency of RE conformational sampling was monitored by measuring $N_{TE}(\tau|T_0)$, the number of round-trip transitions in the conformational state of a replica, conditional on the temperature of interest T_0, that occur in a given observation time τ. A transition event is a transit of a given replica from one conformation at T_0 to the other conformation at T_0 *and back again* regardless of route, i.e., whether it was the result of a direct barrier crossing at T_0 or indirectly via a barrier crossing at some other temperature combined with temperature exchanges. Conceptually, this measure reflects the potential of RE to achieve rapid equilibration at the temperature of interest by means of conformational transitions at temperatures other than the temperature of interest. The transition events as defined correspond to the "reversible folding" events studied in all-atom simulations of peptide systems.[48,49] We will use the symbol N_{TE} as a shorthand notation for $N_{TE}(\tau|T_0)$, where T_0 will generally be the lowest temperature in the simulation. For an uncoupled simulation, the number of transition events is simply the number of round trips between macrostates.

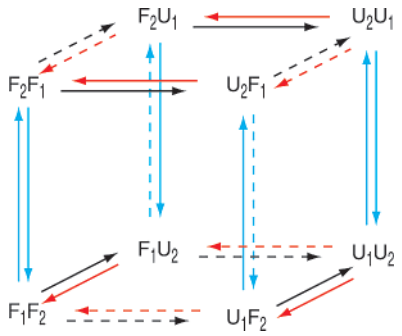

Figure 2. The kinetic network model for the discrete NRE model used by Zheng, et al.[20] The state labels represent the conformation (letter) and temperature (subscript) for each replica. For example, F_2U_1 represents the state in which replica 1 is folded and at temperature T_2, while replica 2 is unfolded and at temperature T_1. Red and black arrows correspond to folding and unfolding transitions, respectively, while the temperature at which the transition occurs is indicated by the solid and dashed lines (for T_2 and T_1, respectively). The cyan arrows correspond to temperature exchange transitions, with the solid and dashed lines denoting transitions with rate parameters α and $w\alpha$, respectively.

2.4. Discrete NRE. We review here the discrete kinetic network model which we devised in our recent study of RE efficiency.[20] In this model (unlike the continuous potential model above), the macrostates F and U (for "folded" and "unfolded") do not possess any internal structure. Instead, it is assumed that the system evolves in time as a Poisson process, in which instantaneous transitions between F and U occur after waiting periods given by exponentially distributed random variables with means equal to the reciprocals of the folding or unfolding rates. The result (for a single replica) is an example of a "random telegraph" Markov process.[50]

If the transition events are Markovian, then the simultaneous behavior of two uncoupled non-interacting replicas can be represented by the four composite states $\{F_1F_2, F_1U_2, U_1F_2, U_1U_2\}$. In each symbol, the first letter represents the configuration of replica 1, the second letter represents the configuration of replica 2, and the subscripts denote the temperature of each replica. Only transitions corresponding to a single conformational change (e.g., $F_1F_2 \rightarrow U_1F_2$) are allowed, assuming that the probability of two simultaneous changes (e.g., $F_1U_2 \rightarrow U_1F_2$) in an infinitesimal interval dt can be neglected.[50] The four-state composite system for two non-interacting replicas can be extended to create a network model of RE by introducing temperature exchanges between replicas, i.e., by allowing transitions such as $F_1U_2 \rightarrow F_2U_1$. This leads to a system with eight states arranged in a cubic network with "horizontal" folding and unfolding transitions and "vertical" temperature exchange transitions (Figure 2). For canonical equilibrium probabilities to be preserved under temperature exchanges, it is sufficient that detailed balance is satisfied, e.g., the transition probabilities $w(F_1U_2 \rightarrow F_2U_1)$ and $w(F_2U_1 \rightarrow F_1U_2)$ satisfy $P_{eq}(F_1U_2)w(F_1U_2 \rightarrow F_2U_1) = P_{eq}(F_2U_1)w(F_2U_1 \rightarrow F_1U_2)$. The ratios of forward and reverse transition probabilities for $F_1F_2 \rightleftharpoons F_2F_1$ and $U_1U_2 \rightleftharpoons U_2U_1$ are equal to 1, as interchange of temperatures does not change the equilibrium populations.

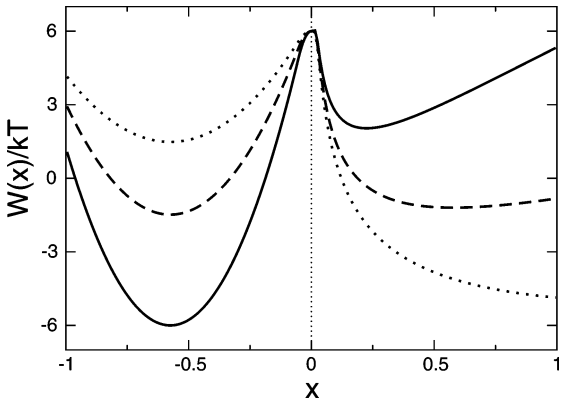

Figure 3. The PMF at three different temperatures: 296 K (solid line), 474 K (dashed line) and 789 K (dotted line). The PMF was calculated using numerical integration. To more clearly illustrate the change in the barrier height as a function of temperature, the three curves have been superimposed to coincide at $x = 0$.

The effect of the rate of temperature exchanges is included by introducing the rate parameter α, which controls the overall scaling of the temperature exchange rate relative to the folding and unfolding rates. The forward and reverse rates of the $F_1F_2 \rightleftharpoons F_2F_1$ and $U_1U_2 \rightleftharpoons U_2U_1$ transitions are set equal to α, while the other rates are set to α or $w\alpha$ as required by detailed balance, where, in this case, $w = P_{eq}(F_2U_1)/P_{eq}(F_1U_2)$ or its reciprocal such that $w < 1$ (see Figure 2). The overall average rate at which temperature exchanges occur (k_{ex}) is the probability of jumping in any instant dt from the upper to the lower face (or vice versa) of the cubic network, and is given by the equilibrium population weighted sum of the temperature exchange rates over all states:

$$k_{ex} = \frac{k_{f1}k_{f2} + 2k_{u1}k_{f2} + k_{u1}k_{u2}}{(k_{f1} + k_{u1})(k_{f2} + k_{u2})}\alpha \qquad (3)$$

The NRE model was simulated using a standard method for continuous time Markov processes with discrete states,[50] also known as the "Gillespie algorithm". Given a current state X_0, we identify its m neighboring states X_1, X_2, ..., X_m and the transition rates k_1, k_2, ..., k_m from X_0 to each of the neighboring states. We generate a waiting time in state X_0 by drawing a random number from an exponential distribution with mean $(k_1 + k_2 + \cdots + k_m)^{-1}$, and select a destination state X_i from among X_1, X_2, ..., X_m with probability $k_i/(k_1 + k_2 + \cdots + k_m)$. This procedure is then repeated with the new state as the current state.

3. Results and Discussion

3.1. Thermodynamics and Kinetics of the Continuous Model System. *3.1.1. Thermodynamics.* In Figure 3 we show the PMFs corresponding to the two-dimensional potential along the x coordinate at several temperatures. PMFs calculated by MC sampling and numerical integration of the canonical distribution function agree to within statistical accuracy. The PMFs show two free energy minima corresponding to the folded (F, $x \leq 0$) and unfolded (U, $x > 0$) conformational states, separated by a free energy barrier near $x = 0$. The free energy minimum of the unfolded state and the free energy barrier have no counterparts in the potential, which is monotonically varying in both of these regions (Figure 1). These features of the PMF originate from the interplay between opposing entropic and enthalpic driving forces. The free energy minimum of the unfolded state corresponds to the optimal balance between entropy, which drives the system toward large values of x (where the accessible space along the y coordinate is greatest), and enthalpy, which drives the system toward small values of x (where the potential energy is smallest). The free energy barrier that separates the unfolded and folded state is entropic in origin. For x near 0, the entropy is significantly reduced compared to the unfolded state, and assumes a value similar to that of the folded state (compare in Figure 1 the size of the accessible space along y at $x = 0$ and for $x > 0$ and $x < 0$). In contrast, the potential energy at $x = 0$, although smaller than in the unfolded state, is still substantially larger than in the folded state. This imbalance between entropy and potential energy causes the free energy maximum at $x = 0$.

From the point of view of folding, the free energy maximum constitutes an entropic bottleneck. In order to make a transition to the folded state, the system needs to cross the free energy barrier region at $x = 0$, where the system has lost all of the entropy required for folding without having gained all of the folding enthalpy. Similar transition bottlenecks have been described in simplified models for protein folding.[34,40,51] After crossing this barrier, the system enters the folded state by going downhill in potential energy without further reduction in conformational entropy, since the accessible space along the y direction is the same for all points x in the folded space. Because the conformational entropy is constant for $x < 0$, the PMF in this region coincides with the potential energy. From the point of view of unfolding, the free energy maximum at $x = 0$ constitutes an enthalpic barrier. Relative to the folded state, points in the region near $x = 0$ have similar conformational entropy but larger potential energy. To reach the barrier region from the folded state, therefore, the system needs to gain potential energy (enthalpy) without the help of a concomitant increase in conformational entropy. Beyond the barrier region there is a free energy gain for moving toward the unfolded state since the gain in conformational entropy outweighs the increase in potential energy.

As shown below, the barrier region close to $x = 0$ constitutes the transition state for the folding/unfolding equilibrium. The free energy difference between the unfolded and folded states and the transition state corresponds to the free energies of activation, which determine the rate of folding and unfolding, respectively. Because of their different thermodynamic origins (entropic vs enthalpic), the free energies of activation for folding and unfolding display the opposite dependence on temperature. As Figure 3 shows, the free energy of activation for folding increases with increasing temperature relative to thermal energy (kT), where the free energy of activation for unfolding decreases with increasing temperature. This anti-Arrhenius behavior is the signature of an entropically activated process. The conformational entropy difference between the unfolded state and the transition state increases as the temperature is increased, leading to an increase in the height of the free energy barrier for folding with increasing temperature.

Figure 4 shows the temperature dependence of the population, $P_F(T)$, of the folded state, often referred to as the melting curve. The shape of the melting curve is typical of two-state protein thermal denaturation experiments. At 300 K, the system is nearly completely folded, and the fraction folded decreases with increasing temperature in favor of the unfolded state which is entropically favored. The melting temperature T_M (corresponding to equal populations of the folded and unfolded state) is approximately 460 K. At this temperature, the folded and

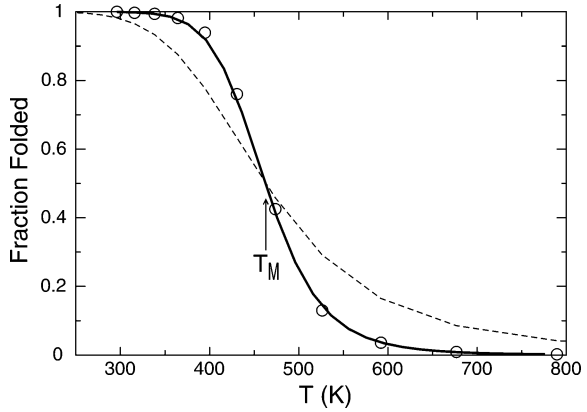

Figure 4. The temperature dependence of the fractional population folded (solid line) calculated by numerical integration of the PMF. The temperature dependence of the fraction folded corresponding to a system with a smaller average potential energy difference between the folded and unfolded states (see Appendix) is shown for comparison (dashed line). The fraction folded derived from the folding and unfolding rates obtained by MC simulation (Figure 6) is shown as circles. The melting temperature $T_M = 463$ K (corresponding to 50% folded population) is indicated.

unfolded states have equal free energy. The slope of the melting curve at the melting temperature is

$$\left(\frac{dP_F}{dT}\right)_{T=T_M} = \frac{1}{4}\frac{\bar{U}_F - \bar{U}_U}{kT^2}$$

which is proportional to the difference of the average potential energies, U_F and U_U, of the folded and unfolded states. Thus, a decrease of the average potential energy difference between the two states leads to a less steep melting curve. To illustrate this, we show in Figure 4 the melting curve corresponding to an alternative parametrization of the potential for which the average potential energy difference between the folded, unfolded, and transition states was decreased, while approximately preserving the same value of the melting temperature (see Appendix). As expected, the alternative parametrization leads to a more gradual conversion from the folded state to the unfolded state with increasing temperature (Figure 4, dashed line). The heat capacity as a function of temperature is approximately Gaussian and is peaked near T_M.

3.1.2. Kinetics. With the MC move set described in the Methods section above, the kinetics of folding/unfolding is two-state as measured by the distribution of first passage times, which is exponential (Figure 5). The Arrhenius plots of the folding and unfolding reaction rates are shown in Figure 6. The temperature dependence of the reaction rates using the Arrhenius equation with activation free energies extracted from the PMFs (Figure 3) agree well with the simulation results, and is a further indication that the kinetics is two-state and that the reaction coordinate is well represented by the x coordinate. This is a consequence of choosing a move set for which equilibration along the y coordinate is faster than that along the x coordinate. The alternative potential parametrization in the Appendix, which is characterized by a smaller average potential energy of the unfolded state relative to the folded and the transition states, leads to a weaker temperature dependence of the folding rate (Figure 6, dashed lines). Since the slope of the Arrhenius curve is proportional to the activation energy, this difference of the rates is consistent with the smaller energy of activation obtained with the alternative parametrization.

The folding rates decrease with increasing temperature, a phenomenon that has been observed in the kinetics of protein

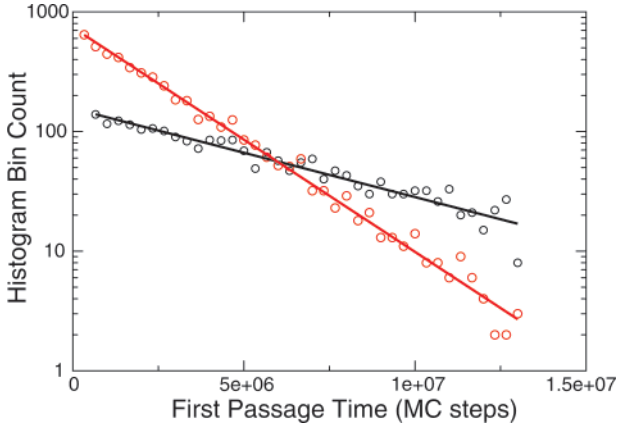

Figure 5. The distributions of first passage times for folding (black) and unfolding (red) observed during a 2.7×10^{10}-step kinetic MC at 475 K. Approximately 4700 folding and unfolding events were observed. A folding first passage time is defined as the time elapsed from when the particle enters the unfolded region from the buffer region (having previously been in the folded region), until it re-enters the folded region. The unfolding first passage time is defined similarly. The semilog plot of the histograms of the first passage times is shown as circles, while the lines represent the best-fit exponential curve.

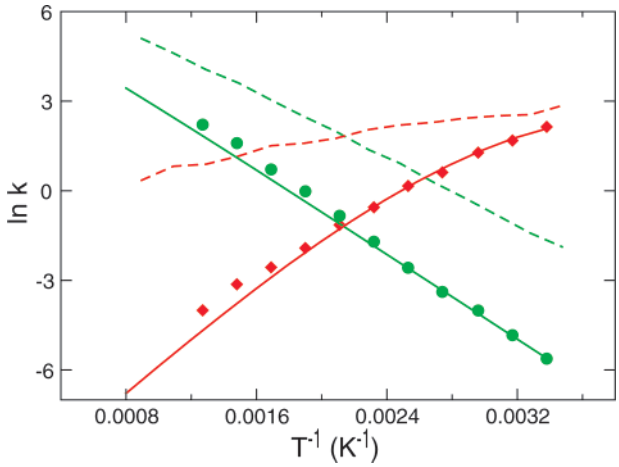

Figure 6. The temperature dependence of the folding and unfolding rate constants (solid lines and symbols). Folding and unfolding rates are indicated by red and green color, respectively. The folding and unfolding rates corresponding to a system with a smaller activation energy for folding (Appendix) are shown for comparison (dashed lines). The rate constants plotted in symbols were derived from kinetic MC simulations run at different temperatures. The solid lines represent the rates calculated using the Arrhenius equation based on activation energies derived from the PMF along x (Figure 3). Rate constants are expressed in units of 10^{-6} per MC step.

folding.[33−36,42] Processes displaying anti-Arrhenius behavior are said to be characterized by a negative effective activation energy, whereby the enthalpy of the unfolded state is larger than that of the transition state. The free energy of activation of these processes, however, remains positive as a result of the activation entropy favoring the unfolded state. The negative activation entropy is associated with the smaller number of accessible conformations at the transition state relative to the unfolded state; that is, the transition state constitutes an entropic "bottleneck'' that needs to be traversed for the transition to the folded state to occur. These elements clearly exist in the simplified two-dimensional system under investigation. Since the potential energy decreases monotonically from the unfolded state to the folded state, the average potential energy at the transition state ($x = 0$) is smaller than the average potential energy of the unfolded state, leading to the observed anti-Arrhenius behavior

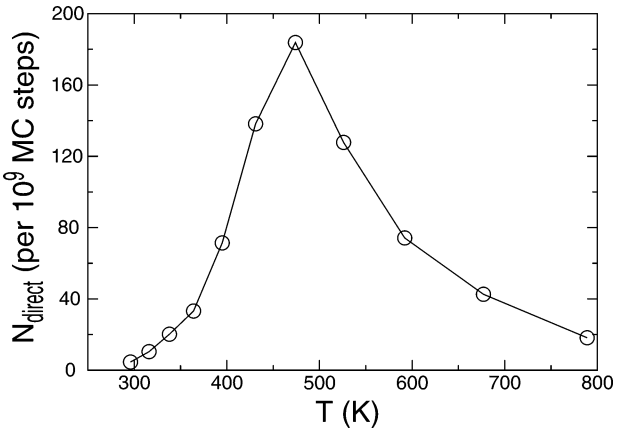

Figure 7. Number of direct round-trip transition events N_{direct} in single-temperature uncoupled simulations over the temperature range 296−789 K in 5×10^9 MC steps. The curve plotted as a solid line was calculated from the harmonic mean of the folding and unfolding rates estimated from the mean of the folding and unfolding first passage time distribution (Figure 5) obtained by MC simulations at each temperature, while the number of events counted directly from the MC simulations at individual temperatures is plotted as circles. The high level of agreement indicates that the system is very well approximated as a two-state activated process.

of the rate of folding. Despite the enthalpic driving force favoring the transition state, the free energy of activation for folding remains positive at all temperatures examined (as the calculated PMF along the *x* coordinate shows). This is because the entropy of the transition state is smaller than the entropy of the unfolded state because of the larger accessible configuration space along the *y* coordinate (Figure 1). The entropic destabilization of the transition state, which (as in protein folding) can be described as acting as a "bottleneck", more than offsets the enthalpic stabilization, leading to the observed positive activation free energy for folding.

Often the observed folding rates of proteins show non-monotonic behavior with respect to the temperature; the folding rate increases with temperature at low temperatures as in normal Arrhenius behavior, switching to anti-Arrhenius behavior at higher temperatures, when the folding rate decreases with increasing temperature. This phenomenon is rationalized in terms of a negative activation heat capacity. The activation heat capacity is defined as the temperature derivative of the activation energy, and a negative value of the activation heat capacity indicates that the unfolded state has a larger heat capacity than the transition state. The observed negative heat capacity of activation of protein folding has been variously interpreted as being due to the hydrophobic effect[33,42] or to the difference of the distribution of energies of the molecular conformations experienced as a function of temperature.[34,52] The curvature of the Arrhenius plot is related to the activation heat capacity. The present simplified two-dimensional system does not have a large enough heat capacity of activation to reproduce this turnover from Arrhenius to anti-Arrhenius behavior within the temperature range we have investigated. Thus, the results extracted from this model are applicable only to the anti-Arrhenius temperature regime of the protein folding process.

Figure 7 shows the number of direct round-trip transition events N_{direct} observed during MC simulations of $N_{MC} = 5 \times 10^9$ steps as a function of temperature. We use the number of transitions as a measure of the efficiency of conformational sampling, which determines the rate of convergence of thermodynamic quantities extracted from the simulations. The results of Figure 7 show that the conformational sampling efficiency of the uncoupled simulation varies non-monotonically with the temperature. There is a 40-fold increase in transitions from 300 to 474 K, the temperature at which the maximum is observed. This decreases for temperatures higher than 474 K, reaching a 10-fold reduction at 800 K (relative to the maximum). As the results in Figure 7 show, this behavior mirrors almost exactly the behavior of the harmonic mean $(k_f^{-1} + k_u^{-1})^{-1}$ of the folding and unfolding rates (from Figure 6) as a function of temperature (we note that our use of the term "harmonic mean" differs from standard usage by a factor of 2, which is natural given that we are considering a round trip, i.e., a single "transition event" involves two conformational transitions). The agreement between the harmonic mean of the rates and the number of direct round trip transitions is expected for a two-state activated equilibrium, since the average time of a round-trip excursion from the folded to the unfolded state and back is the sum of the average folding and unfolding times $\tau_f = k_f^{-1}$ and $\tau_u = k_u^{-1}$, respectively: $N_{direct} = N_{MC}/(\tau_f + \tau_u)$.

3.2. RE Simulations Using MC on the Continuous Potential. In a recent paper,[20] we analyzed the convergence and efficiency of RE using a discrete model for folding and unfolding. We found that when the physical kinetics shows anti-Arrhenius temperature dependence, there exists an optimal maximal temperature beyond which the efficiency of the RE method is degraded. Similar behavior is expected from RE simulations using the continuous two-dimensional potential, with possible differences arising from the more complex nature of the present model, where the folded and unfolded states have internal structure. We performed RE simulations on the continuous two-dimensional potential with MC as the dynamic propagator and RE proposals made periodically between adjacent temperatures every N_X MC steps. The efficiency of conformational sampling was monitored by counting the number of temperature-conditional transition events N_{TE} defined in section 2.3 above.

In order to directly compare with the results obtained previously, we first performed RE using two replicas. Although such a simulation would not be realistic in general for a protein system due to poor energy overlap and very inefficient temperature exchange, it is feasible in the two-dimensional potential. The result for a 2×10^9-step simulation where the lower temperature is held fixed at 296 K and the upper temperature varies from 296 to 789 K is shown in Figure 8 (green, red, and blue dots). We see behavior similar to that seen for the discrete model studied previously: the number of temperature-conditional transitions N_{TE} has non-monotonic behavior and exhibits a maximum at an optimal high temperature given by the maximal harmonic mean of the folding and unfolding rates (474 K). This maximum point is approximately independent of the rate at which attempted temperature exchanges occur. While the location of the maximum is in agreement with our previous results,[20] the magnitude of the number of transition events is not. We have shown that, for NRE simulations employing a two-state model (folded and unfolded states), the number of transition events is given by the average over all temperatures of the harmonic means of the folding and unfolding rates, provided that the rate of temperature exchanges is sufficiently fast.[20] In the continuous model, we find that the number of transitions is significantly lower than that predicted from the average of the harmonic means of the rates (Figure 8, black dashed line). This may be due to the finite rate of temperature exchanges, deviations from the pure Markovian kinetics of the two-state discrete model, or a combination of these effects.

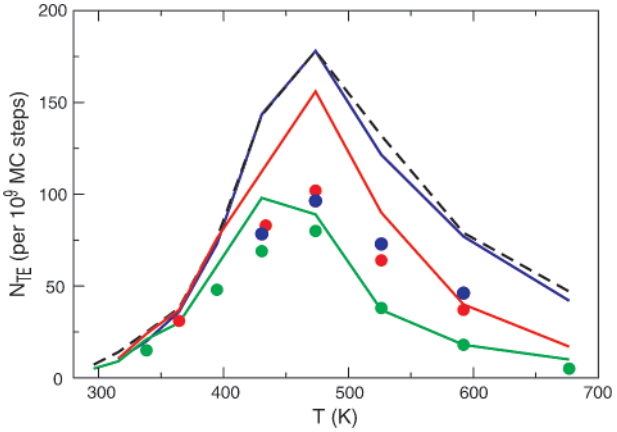

Figure 8. The dependence of the number of temperature-conditional transition events N_{TE} (section 2.3) on the temperature of the high-temperature replica for a two-replica simulation on the continuous potential (circles), and comparison with predicted transition events using the discrete NRE model (Figure 2) (curves). For all simulations, one replica was fixed at 296 K and the other replica was scanned from 296 to 789 K. The black dashed line corresponds to the discrete model prediction in the large-α limit. The solid curves are the predicted N_{TE} using the NRE model with temperature-dependent folding and unfolding rates taken from the kinetic MC simulations on the continuous potential (shown in Figure 6) and using an α corresponding to an attempted temperature exchange. The circles are the N_{TE} values observed in 2×10^9 MC step RE simulations on the continuous potential. The green, red, and blue data correspond to N_X values of 1 000, 200, and 20, respectively.

TABLE 1: Number of Temperature-Conditional Transition Events in 2×10^9 MC Steps for Two Replicas[a] as a Function of the Number of MC Steps between Attempted Temperature Exchanges (N_X), and Observed Temperature-Conditional MFPTs[b]

	N_{TE} per replica		temperature-conditional MFPTs	
N_X	observed (continuous)	predicted (NRE)	$F_1 \rightarrow U_1$	$U_1 \rightarrow F_1$
10 000	22	24	91.8	5.6
2 000	52	73	31.6	5.7
1 000	80	105	23.1	5.7
500	93	134	19.1	5.8
200	102	162	16.0	5.7
100	99	168	14.3	5.9
80	98	172	14.6	5.8
50	98	176	14.8	5.7
20	96	177	14.9	6.1
0[c]		178		

[a] With temperatures of 296 and 474 K. [b] In units of 10^6 MC steps; see text for details. [c] Predicted N_{TE} based on the harmonic mean relationship for the $\alpha \rightarrow \infty$ limit.

To test whether this reduced number of transitions is due to insufficiently fast temperature exchange attempts, we performed several simulations in which we varied N_X (the number of MC steps between attempted temperature exchanges). We see in Table 1 that N_{TE} is approximately constant provided that the attempted exchange rate is faster than a critical value of $N_X \approx 500$. For less frequent exchange attempts, we see a substantial decrease in the number of transitions. Thus, the number of unfolding and refolding transitions cannot be increased simply by increasing the rate of attempted exchanges.

3.3. Non-Markovian Effects Revealed by Comparison of Continuous and Discrete RE Simulations. To explore causes for the observed transition deficit, we performed simulations using the discrete NRE model (Figure 2) using kinetic parameters derived from the two-dimensional continuous potential (Figure 6). To map the rates determined using the continuous potential to the discrete model, we used the folding and unfolding rates directly, expressed in units of 10^{-6} per MC step. Different values of α were used for the $F_1F_2 \rightleftharpoons F_2F_1$ and $U_1U_2 \rightleftharpoons U_2U_1$, and were set to $10^6/N_X$ multiplied by the empirical acceptance rate when both replicas are in the folded or unfolded state (0.853 and 0.395, respectively).

TABLE 2: Empirical "Reverse-Engineered" Rates at Temperatures $T_1 = 296$ K and $T_2 = 474$ K (in Units of 10^{-6} MC Step) from Continuous Potential Simulation Data Assuming the Network Topology of Figure 2

	uncoupled rates	reverse-engineered rates			
		$N_X = 10\,000$	$N_X = 2\,000$	$N_X = 200$	$N_X = 100$
k_{f1}	6.08	5.66	6.10	5.27	6.33
k_{u1}	0.0036	0.0038	0.0036	0.0037	0.0037
k_{f2}	0.297	0.288	0.299	0.290	0.306
k_{u2}	0.420	0.420	0.419	0.427	0.425

If we compare the observed number of transitions seen in the continuous model with the number predicted by the NRE model with the same rate parameters (Table 1), we see that there is good agreement when the attempted exchange rate is small, but substantial disagreement when it becomes larger. In particular, while the number of transitions using the continuous model reaches a plateau value at $N_X \approx 1000$, the predicted number of transitions in the NRE model continues to increase, asymptotically approaching the value predicted by the average of harmonic means. Similarly, comparison of the predicted and observed number of transitions as a function of temperature (Figure 8) shows a significant overestimation of the transition rate by the NRE model, and shows that this overestimation is much more severe when the rate of attempted temperature exchanges is fast. For example, while the N_{TE} predicted from the NRE model has essentially reached the asymptotic limit when $N_X = 20$ (blue curve), the observed N_{TE} values are essentially unchanged relative to those obtained when $N_X = 200$ (compare blue and red circles). The continuous two-dimensional model thus appears to contain an inherent "speed limit", which prevents it from achieving the transition rates expected for a fully Markovian system, even if the temperature exchanges are attempted frequently.

One possible origin of this speed limit is that the average effective rates are different in the coupled and uncoupled systems. To test this, we analyzed the kinetics of the continuous RE simulation by using the NRE model to "reverse-engineer" the apparent rates by estimating the mean residence times and branching ratios for various RE macrostates. If the system is Markovian, then the rate k_{tot} given by the inverse of the mean residence time is the sum of the rates exiting that state. The rate corresponding to a given edge can then be estimated by multiplying k_{tot} by the fraction of residences that exit via that edge (the branching ratio). The results are shown in Table 2. The reverse-engineered rates generally agree with the uncoupled folding and unfolding rates estimated from kinetic MC, and this is true for both rapid and slow attempted temperature exchange rates. Therefore, the temperature exchanges do not perturb the average kinetics of the system, and cannot be a cause of the limit on the transition rates at rapid temperature exchange rates.

In order to further investigate the origin of the observed speed limit, we calculated the MFPTs for temperature conditional folding and unfolding, i.e., the average time for a replica unfolded at low temperature to become folded at low temperature (regardless of path), or vice versa. The resulting MFPTs for the continuous potential are shown in Table 1. We see there that the N_{TE} speed limit arises exclusively from a limitation in the fastest achievable unfolding rate, since the folding process

is independent of N_X and is not rate limiting. This can be understood by noting that the values of α corresponding to the N_X values used are at least 2 orders of magnitude larger than the folding and unfolding rates. To unfold, the system need only make use of temperature exchange transitions that correspond to α (i.e., the solid cyan arrows of Figure 2). Since α is already much larger than the other rates, changes to it due to changes in N_X will not significantly change the MFPT for folding.

On the other hand, the unfolding process (if it occurs via an indirect route, which is likely given the very small value of k_{u1}) requires the system to use a "$w\alpha$ edge" (i.e., a dashed cyan arrow in Figure 2). Since $w \approx 10^{-4}$ for the temperatures used here, $w\alpha$ is now slower than or comparable to the folding and unfolding rates, and therefore changes in N_X can make a substantial impact on the unfolding MFPT. Thus, the N_{TE} speed limit can be traced to the kinetics of temperature conditional unfolding, and must arise from some difference between the unfolding kinetics in the continuous potential and the fully Markovian NRE model.

One obvious way in which the continuous and NRE models differ is that the macrostates in the continuous potential have spatial extent, unlike the NRE states which lack internal structure. This means that a finite time is required for the particle to transit the nonequivalent microstates that make up the two wells. In fact, we observe that the correlation time for diffusion in the *x* direction *in the unfolded well* at 474 K is approximately 1 400 MC steps. This time scale is of the same magnitude as the N_X value at which the speed limit effect of Table 1 begins to occur, suggesting that there may in fact be a connection between the observed N_{TE} speed limit and conformational diffusion within the free energy wells. Such dependence of the kinetics on the internal structure of the macrostate can lead to non-Markovian behavior.

Formally, a process is Markovian if and only if the observed propagators (Green's functions) do not depend on the history of the trajectory prior to the current state, i.e.,

$$P(x_3,t_3|x_1,t_1;x_2,t_2) = P(x_3,t_3|x_2,t_2) \tag{4}$$

for all states x_1, x_2, x_3 and all times $t_1 < t_2 < t_3$. Although eq 4 could be used to directly detect deviations from Markovian behavior, previous work has typically used other analysis methods to detect such deviations.[29,53,54] For example, in a Markovian process, the rate matrix **K** determines the propagators via the master equation

$$\dot{\mathbf{p}}(t) = \mathbf{K}\mathbf{p}(t) \tag{5}$$

where $\mathbf{p}(t)$ is the vector of propagators at time *t*. The formal solution of eq 5 is given by $\mathbf{p}(t) = e^{\mathbf{K}t}\mathbf{p}(0)$, and therefore $e^{\mathbf{K}\tau}$ can be thought of as a transition matrix $\mathbf{T}(\tau)$, i.e., the matrix of probabilities of being in state x_j at time τ given that the system was in state x_i at time 0. If we denote the eigenvalues of **K** by $\lambda_1 > \lambda_2 > \cdots$ and the eigenvalues of $\mathbf{T}(\tau)$ by $\mu_1(\tau) > \mu_2(\tau) > \cdots$, then $\mu_i(\tau) = e^{\lambda_i\tau}$. This can be used as a test of Markovian behavior, since $\mathbf{T}(\tau)$ can be empirically estimated from a trajectory. Different values of the lag time τ will yield different values of $\mu_i(\tau)$; however, $\tau/\ln \mu_i(\tau)$ should be independent of τ if the kinetics is Markovian.[29,54] Alternatively, the Markov property can be tested by analyzing the transition probabilities as a function of lag time using an information theoretic measure based on Shannon's entropy.[53]

We have chosen to detect deviations from Markovian kinetics by examining the observed residence time distributions and branching ratios, which provides insights into the physical origin and the mechanism by which the non-Markovian effects enter into the stochastic process. In our simulations on the continuous potential, we have found that the residence time distributions in the macrostates are exponential to within statistical uncertainty (data not shown), and thus by themselves are consistent with Markovian kinetics. The branching probabilities, however, are significantly dependent on the preceding macrostate. We focused on transitions entering and leaving the thermodynamically favored U_2F_1 macrostate (or its symmetry-related state F_1U_2). We ran several trajectories using different rates of attempted temperature exchange and tallied the number of times each macrostate sequence (X,U_2F_1,Y) was observed in each (where $X,Y \in \{F_2F_1,U_2U_1,U_1F_2\}$). These counts were transformed into normalized branching probabilities, where $P(X|Y)$ denotes the history-independent branching probability of next visiting macrostate X given that the system is currently in macrostate Y, and $P(X|Z,Y)$ denotes the history-dependent branching probability of next visiting macrostate X given that the system is currently in macrostate Y and had been in macrostate Z immediately prior (Table 3).

TABLE 3: History Dependent and Independent Branching Probabilities from State U_2F_1

	maximum likelihood estimate[a] (and 95% credible interval[b])	
conditional probability	$N_X = 200$	$N_X = 10\,000$
$P(U_1F_2\|U_1F_2,U_2F_1)$	0.906 (0.904, 0.908)	0.168 (0.144, 0.195)
$P(U_1F_2\|F_1F_2,U_2F_1)$	0.521 (0.153, 0.530)	0.094 (0.088, 0.101)
$P(U_1F_2\|U_2F_1)$	0.849 (0.846, 0.851)	0.103 (0.096, 0.110)
$P(F_2F_1\|F_2F_1,U_2F_1)$	0.477 (0.469, 0.486)	0.895 (0.888, 0.902)
$P(F_2F_1\|U_1F_2,U_2F_1)$	0.092 (0.090, 0.094)	0.816 (0.788, 0.841)
$P(F_2F_1\|U_2F_1)$	0.150 (0.147, 0.152)	0.886 (0.878, 0.893)

[a] Maximum likelihood estimates determined using $P(a_1|b,c) = \#(b,c,a_1)/\sum_i \#(b,c,a_i)$ and $P(a_1|b) = \sum_i \#(c_i,b,a_1)/\sum_{jk} \#(c_j,b,a_k)$, where $\#(i,j,k)$ is the number of occurrences of the ordered triple (i,j,k). [b] Bayesian credible intervals under a uniform prior given by the 0.025 and 0.975 quantiles of the distribution $P(p) \propto p^n(1-p)^{N-n}$, where n and N are the numerator and denominator, respectively, of the fraction used to calculate the maximum likelihood estimate.

If the kinetics is Markovian, then the history-dependent and corresponding history-independent branching probabilities will be equal:

$$P(X|Z,Y) = P(X|Y)$$

from which it follows that history-dependent branching probabilities that differ only in the history condition will also be equal:

$$P(X|Z,Y) = P(X|W,Y)$$

This is clearly not the case for the data in Table 3. For example, we see that the history-dependent branching probabilities $P(U_1F_2|F_2F_1,U_2F_1)$ and $P(F_2F_1|F_2F_1,U_2F_1)$ differ significantly from their corresponding history-independent branching probabilities $P(U_1F_2|U_2F_1)$ and $P(F_2F_1|U_2F_1)$, and the branching probability $P(U_1F_2|F_2F_1,U_2F_1)$ is significantly smaller than $P(U_1F_2|U_1F_2,U_2F_1)$. This is most pronounced when the rate of attempted temperature exchanges is fast.

Examination of the kinetic scheme of Figure 2 indicates that the deviations from Markovian behavior seen in Table 3 are consistent with a reduction in the number of temperature-conditional round-trip conformational transition events. If the unfolding rate at low temperature is negligible, then a low-temperature folded conformation unfolds predominantly via indirect paths of the form $F_1F_2 \rightarrow F_2F_1 \rightarrow U_2F_1 \rightarrow U_1F_2$ or $F_1U_2 \rightarrow F_2U_1 \rightarrow U_2U_1 \rightarrow U_1U_2$. In the former case, the F_2F_1

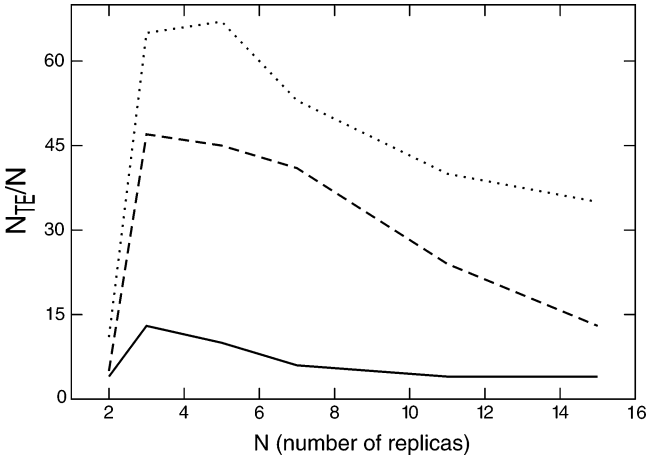

Figure 9. Number of transition events N_{TE} (section 2.3) normalized by the number of replicas in 2×10^9 MC steps for 2 to 15 replicas exponentially distributed in temperature from 296 to 789 K. Temperature exchanges were attempted every 10 000 (solid line), 1 000 (dashed line), and 200 (dotted line) MC steps.

$\rightarrow U_2F_1$ step is more likely to be reversed when the temperature exchange rate is rapid (Table 3), as is the $F_1U_2 \rightarrow F_2U_1$ step in the latter case (which follows by symmetry from the $U_2F_1 \rightarrow U_1F_2$ results of Table 3). Thus, increasing the rate of attempted temperature exchanges increases the probability of counterproductive backtracking relative to the Markovian case, resulting in a decrease in the rate of temperature-conditional unfolding events, and therefore a corresponding decrease in N_{TE} (since temperature-conditional unfolding was shown above to be rate-limiting).

Although the results presented here do not identify the physical origin of the non-Markovian kinetics, we hypothesize that it is due to the finite time required for diffusion of the particle within the macrostates. This effect does not arise in the NRE model, since, in there, the macrostates have no internal structure, and the probability of making a transition to a given macrostate at any instant *dt* is the same, regardless of which macrostate the system was in previously or how long it has been in the current macrostate. The behavior of the continuous system within the wells is not Markovian, since the system has memory that is mediated by conformational diffusion within the macrostate. This correlation in time of the particle's position (and energy) implies that there is a maximal effective value of the rate of statistically independent temperature exchanges, which is limited by the time required for conformational relaxation *within* the folded and unfolded macrostates.

3.4. Dependence of RE Efficiency on the Number of Replicas. The above results were obtained with two replicas, which is not typical for RE simulations that would be carried out for peptides and proteins. To investigate the effect of adding additional replicas, we performed a series of simulations of 2×10^9 MC steps with 2 to 15 replicas distributed uniformly in T^{-1} from 296 to 789 K. The results are shown in Figure 9. One important issue that arises when considering such a set of results is the appropriate measure of conformational sampling efficiency of RE. If we consider the total number of transition events N_{TE} (direct and indirect) in all replicas, then we would see for the most part a monotonic increase of efficiency as a function of the number of replicas *N* simply because the number of indirect "channels" for transitions is linearly increasing. This measure of efficiency, however, implicitly assumes that computer power is inexpensive and that the convergence rate of the simulation is the important limiting factor. If both computer resources and the convergence rate are limiting factors, a more appropriate

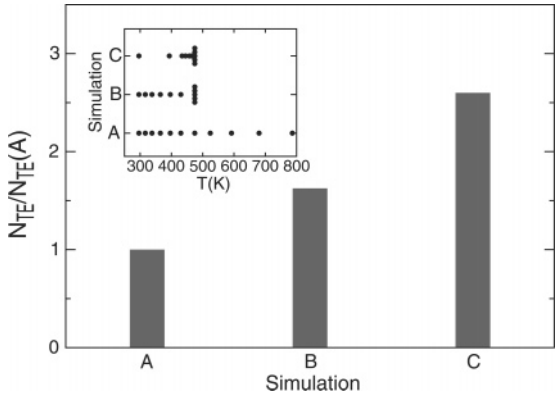

Figure 10. Number of transition events N_{TE} (section 2.3) observed in 2×10^9 MC steps for three different 11-replica RE simulations performed using the continuous potential with $N_X = 200$. The temperature distributions for the three simulations are shown in the inset. Transition counts have been normalized by the N_{TE} of simulation A.

measure is the computational efficiency calculated as the number of transition events per replica (N_{TE}/N). According to this measure, a RE simulation with $N + 1$ replicas is considered more efficient than one with *N* replicas only if the introduction of the additional replica provides more than a proportional increase in the number of transition events at the temperature of interest.

We find that the efficiency increases strongly as a function of *N* when *N* is small, reaches a maximum, and then decreases with *N* for larger *N* (Figure 9). This pattern is unchanged as a function of the rate of attempted temperature exchanges, showing a scaling approximately consistent with the results in Table 1. The trends seen here are qualitatively similar to that seen previously in the NRE two-state discrete model[20] with finite α. In that work, we attributed the decrease with increasing number of replicas beyond an optimum value in part to a combinatoric effect that decreases the relative size of the "target" space of configurations in which a replica is at the temperature of interest relative to the total temperature/configuration space. It is reasonable to assume that a similar effect is occurring here as well. We will address this in a future communication.

The results in Figure 9 were obtained with a relatively uniform distribution of temperatures. It is of interest to consider the effect on efficiency of changing that temperature distribution. In our previous work,[20] we concluded that, in the context of the discrete network model in the "large α" limit, the optimal temperature distribution is one replica at the temperature of interest, and the rest at the temperature that maximizes the harmonic mean of the folding and unfolding rates. That model, however, was limited in its realism in that it did not have explicit energy distribution functions. Furthermore, it is clear from the results presented in the previous section that very large effective values of α may not be achievable in real systems. The continuous two-dimensional potential studied here provides a better test system for studying these questions.

In Figure 10 we show the relative number of temperature-conditional transition events in 2×10^9 MC steps for three different temperature distributions of 11 replicas: (A) uniformly distributed in T^{-1} from 296 to 789 K, (B) 6 replicas uniformly distributed in T^{-1} from 296 to 474 K (the optimal temperature) and the remaining 5 "bunched up" at the optimal temperature, and (C) 5 replicas bunched up at the optimal temperature with the remaining distributed in the 296 to 474 K range but strongly skewed toward the optimal temperature. Temperature distribution B provides more than a 50% increase in efficiency relative

to the uniform distribution over the large temperature range. This is consistent with our discrete model results, and indicates that it is possible to include temperatures that are "too high" when the system exhibits anti-Arrhenius kinetics. However, we can increase the efficiency even further (to more than a factor of 2.5 over the baseline result) by skewing the temperature distribution to increase the number of replicas in the vicinity of the transition temperature (distribution C). Previous work by Hansmann et al. has suggested that such concentration of the temperatures near a bottleneck can improve temperature mixing.[26] However, the improved efficiency may simply be due to the increased number of replicas near the optimal temperature. The clarification of the relative contributions from these two effects will also be addressed in a future communication.

4. Conclusions

One of the challenges of studying the computational efficiency of RE has been the difficulty in running molecular simulations sufficiently long to obtain full convergence and meaningful statistics. This is particularly daunting if such simulations must be run multiple times to assess the effect of differences in simulation protocols and parameters. The use of simplified model systems allows for thorough theoretical, conceptual, and computational analysis of the problem that can provide insights into the factors that limit the efficiency of RE in more realistic molecular systems.

Our previous work made use of a highly simplified discrete model for protein folding with two conformational states at several temperatures.[20] While this system did provide useful insights, it was limited in a number of ways, and, in particular, was fully Markovian. Here we have described a two-dimensional continuous potential function and an associated move set that allows us to perform MC and RE MC simulations in a system that is small enough to quickly converge but yet is rich in a complexity that is reminiscent of molecular systems. While many of the results are consistent with those observed previously, novel effects are also seen. In particular, we have confirmed that the efficiency of RE in more complex systems is fundamentally limited by the time scale of conformational diffusion within basins, as we had anticipated.[20] We expect that such behavior will also be present (perhaps even more strongly) in molecular systems.

There are many unresolved questions raised by this work. One question for which our two-dimensional system would be a good model is for studying the relationship between conformational and thermal diffusion. Optimization of the diffusion of replicas in temperature space has been a major focus of recent theoretical and computational study of the RE method.[3,21−27] However, the convergence of thermodynamic quantities is not limited by thermal diffusion *per se*, but by the exploration of the conformational space of the system. While very poor thermal diffusion obviously defeats the purpose of RE by effectively reducing it to a set of parallel uncoupled simulations, it is not clear that further optimization of thermal diffusion that is already "reasonably good" will automatically improve convergence. Some recent work has begun to address the role of basin-to-basin transitions.[18,28] Similarly, some work on the optimization of thermal diffusion has emphasized the role of temperature bottlenecks,[26] which may turn out to be fundamentally conformational in nature. The exact relationship between thermal and conformational diffusion remains to be fully clarified, and we look forward to studying this and other questions using simplified continuous and discrete models of RE.

Acknowledgment. We thank Attila Szabo for discussions concerning this work. R.M.L. wishes to express his pleasure at being fortunate enough to have known Attila for 30 years and his admiration for Attila's intellectual honesty and style of doing science. We are very glad to participate in this special issue of the Journal of Physical Chemistry which honors Attila Szabo on his 60th birthday. This work has been supported by a grant from the National Institutes of Health (GM30580).

5. Appendix

The alternative potential with decreased average potential energy differences between folded, unfolded, and transition states is of the same general form as the primary potential described in the Methods section and Figure 1, but with the boundary function parameters $\delta = 10^{-5}$, $b = 1$, and $n_1 = 3.5$ and potential energy

$$U(x,y) = \begin{cases} a_1(x + x_0)^2 + b_1y^2 & -1 \le x < -x_1 \quad 0 \le y \le B(x) \\ -a_2x^2 + b_1y^2 + c_0 & -x_1 \le x \le 0 \quad 0 \le y \le B(x) \\ a_3x^2 + b_1y^2 + c_0 & 0 < x \le 1, \quad 0 \le y \le B(x) \\ \infty & \text{otherwise} \end{cases}$$

with $a_1 = 25$ kcal/mol, $a_2 = 250$ kcal/mol, $a_3 = 10$ kcal/mol, $b_1 = 1000$ kcal/mol, and $c_0 = 6$ kcal/mol. The constants x_0 and x_1 were the same as for the primary potential.

References and Notes

(1) Swendsen, R. H.; Wang, J.-S. *Phys. Rev. Lett.* **1986**, *57*, 2607−2609.
(2) Geyer, C. J.; Thompson, E. A. *J. Am. Stat. Assoc.* **1995**, *90*, 909−920.
(3) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604−1608.
(4) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140−150.
(5) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141−151.
(6) Rhee, Y. M.; Pande, V. S. *Biophys. J.* **2003**, *84*, 775−786.
(7) Nymeyer, H.; Gnanakaran, S.; García, A. E. *Meth. Enzymol.* **2004**, *383*, 119−149.
(8) Cecchini, M.; Rao, F.; Seeber, M.; Caflisch, A. *J. Chem. Phys.* **2004**, *121*, 10748.
(9) Tsai, H.-H. G.; Reches, M.; Tsai, C.-J.; Gunasekaran, K.; Gazit, E.; Nussinov, R. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 8174−8179.
(10) Baumketner, A.; Shea, J.-E. *Biophys. J.* **2005**, *89*, 1493−1503.
(11) Verkhivker, G. M.; Rejto, P. A.; Bouzida, D.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Gehlhaar, D. K.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. *Chem. Phys. Lett.* **2001**, *337*, 181−189.
(12) Ravindranathan, K. P.; Gallicchio, E.; Friesner, R. A.; McDermott, A. E.; Levy, R. M. *J. Am. Chem. Soc.* **2006**, *128*, 5786−5791.
(13) Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113*, 6042−6051.
(14) Fukunishi, H.; Watanabe, O.; Takada, S. *J. Chem. Phys.* **2002**, 116, 9058−9067.
(15) Kwak, W.; Hansmann, U. H. E. *Phys. Rev. Lett.* **2005**, *95*, 138102.
(16) Liu, P.; Huang, X.; Zhou, R.; Berne, B. J. *J. Phys. Chem. B* **2006**, *110*, 19018−19022.
(17) Min, D.; Li, H.; Li, G.; Bitetti-Putzer, R.; Yang, W. Dual-topology Hamiltonian-replica-exchange overlap histogramming method to calculate relative free energy difference in rough energy landscape. http://arxiv.org/abs/physics/0605005v1, accessed 10/07.
(18) Zuckerman, D. M.; Lyman, E. *J. Chem. Theory Comput.* **2006**, *2*, 1200−1202.
(19) Beck, D. A. C.; White, G. W. N.; Daggett, V. *J. Struct. Biol.* **2007**, *157*, 514−523.
(20) Zheng, W.; Andrec, M.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 15340−15345.
(21) Kofke, D. A. *J. Chem. Phys.* **2002**, *117*, 6911−6914.
(22) Kone, A.; Kofke, D. A. *J. Chem. Phys.* **2005**, *122*, 206101.
(23) Predescu, C.; Predescu, M.; Ciobanu, C. V. *J. Chem. Phys.* **2004**, *120*, 4119−4128.
(24) Predescu, C.; Predescu, M.; Ciobanu, C. V. *J. Phys. Chem. B* **2005**, *109*, 4189−4196.
(25) Rathore, N.; Chopra, M.; de Pablo, J. J. *J. Chem. Phys.* **2005**, 122, 024111.
(26) Trebst, S.; Troyer, M.; Hansmann, U. H. E. *J. Chem. Phys.* **2006**, *124*, 174903.

(27) Nadler, W.; Hansmann, U. H. E. On dynamics and optical number of replicas in parallel tempering simulations. http://arxiv.org/abs/0709.3289v1, accessed 10/07.
(28) Zuckerman, D. M. *J. Chem. Theory Comput.* **2006**, *2*, 1693.
(29) Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, 108, 6571−6581.
(30) 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.
(31) Andrec, M.; Felts, A. K.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6801−6806.
(32) van der Spoel, D.; Seibert, M. M. *Phys. Rev. Lett.* **2006**, *96*, 238102.
(33) Oliveberg, M.; Tan, Y.-J.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 8926−8929.
(34) Karplus, M. *J. Phys. Chem. B* **2000**, *104*, 11−27.
(35) Ferrara, P.; Apostolakis, J.; Caflisch, A. *J. Phys. Chem. B* **2000**, *104*, 5000−5010.
(36) Yang, W. Y.; Gruebele, M. *Biochemistry* **2004**, *43*, 13018−13025.
(37) Zhang, W.; Wu, C.; Duan, Y. *J. Chem. Phys.* **2005**, *123*, 154105.
(38) Periole, X.; Mark, A. E. *J. Chem. Phys.* **2007**, *126*, 014903.
(39) Gallicchio, E.; Andrec, M.; Felts, A. K.; Levy, R. M. *J. Phys. Chem. B* **2005**, *109*, 6722−6731.
(40) Bicout, D.; Szabo, A. *Protein Sci.* **2000**, *9*, 452−465.
(41) Zhou, H.-X.; Zwanzig, R. *J. Chem. Phys.* **1991**, *94*, 6147−6151.
(42) Chan, H. S.; Dill, K. A. *Proteins* **1998**, *30*, 2−33.
(43) Kolinski, A.; Skolnick, J. *Polymer* **2004**, *45*, 511−524.
(44) Tiana, G.; Sutto, L.; Broglia, R. A. *Physica A* **2007**, *380*, 241−249.
(45) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: Oxford, 1987.
(46) Chandler, D. *J. Chem. Phys.* **1978**, *68*, 2959−2970.
(47) Levy, R. M.; Karplus, M.; McCammon, J. A. *Chem. Phys. Lett.* **1979**, *65*, 4−11.
(48) Rao, F.; Caflisch, A. *J. Chem. Phys.* **2003**, *119*, 4035−4042.
(49) Seibert, M. M.; Patriksson, A.; Hess, B.; van der Spoel, D. *J. Mol. Biol.* **2005**, *354*, 173−183.
(50) Gillespie, D. T. *Markov Processes: An Introduction for Physical Scientists*; Academic Press: Boston, 1992.
(51) Borreguero, J.; Dokholyan, N.; Buldyrev, S.; Shakhnovich, E.; Stanley, H. *J. Mol. Biol.* **2002**, *318*, 863−876.
(52) Bryngelson, J. D.; Wolynes, P. G. *J. Phys. Chem.* **1989**, *93*, 6902−6915.
(53) Park, S.; Pande, V. S. *J. Chem. Phys.* **2006**, *124*, 054118.
(54) Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. *J. Chem. Phys.* **2007**, *126*, 155102.

Chapter 4

Recovering Folding Kinetics From Replica Exchange Simulations With a Kinetic Network Calibrated Using Local Dynamics

4.1 Introduction

Protein folding is a fundamental problem in modern structural biology, 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[41, 42, 43].

If the process in question is activated, 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 an MC procedure is used to sample entire time-ordered paths connecting reactant and product wells in a well-defined manner [44]. 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 transition states.

A popular alternative takes advantage of heterogeneous distributed computing [45, 46] to enhance sampling by combining information from a large number of short molecular dynamics (MD) trajectories steered by rare events ("Folding@Home"). In a similar spirit,

the “milestoning” technique makes use of many short simulations that span two predefined critical points along a given reaction path[47]. While both approaches are powerful strategies, the former can introduce a bias towards fast events in the ensemble average of the reactive trajectories [48], while the latter is limited to a single reaction path that must be specified in advance.

A related set of methods for obtaining kinetic information is based on the use of stochastic dynamics on a free energy landscape [49, 50, 51, 52, 53, 54]. 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[55, 56, 57], 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 multiple transition states.

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[58], and the kinetics on this graph is often assumed to have Markovian behavior[59, 60, 61, 62, 63]. This approach is particularly well suited for reduced lattice models, and was first introduced in that context[59]. For systems with a continuous state space, some form of discretization is required. This can be done by clustering based on chosen reduced coordinates[58, 61], though the clusters must be chosen carefully so as to satisfy the Markovian condition[62, 63, 64, 65]. Alternatively, the discretization can be based on an analysis of the minima and/or saddle points of the energy surface[60, 66, 67], 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[68]. 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[67, 69]. A hybrid approach has also been proposed that makes use of molecular dynamics to infer local transition regions to build disconnectivity graphs[70].

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[51, 56], 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.

Powerful generalized ensemble methods[71] such as replica exchange molecular dynamics (REMD) [72] have been developed which enhance the ability to obtain accurate canonical populations in complex systems by increasing sampling efficiency. However, since REMD involves temperature swaps between MD trajectories, it is not straightforward to obtain kinetic information from such simulations.[63, 73, 54]. Our laboratory has made use of a kinetic network model[74] in which the nodes correspond to molecular conformations from REMD simulation trajectories, and the edges are derived from an ansatz based on structural similarity. While this model was shown to yield physically plausible kinetics[74], 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 multi-dimensional configurational space of the system by running

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)[75], 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 recursively until agreement is obtained between the drift velocities and diffusion coefficients derived from simulations on the network with 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 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 is constructed based on virtually all degrees of freedom, this allows for multiple pathways and 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.

4.2 Methods

4.2.1 Kinetics of the two-dimensional potential and the representation of drift velocity and diffusion coefficient

We use a two-dimensional potential (Fig. 4.1) constructed to mimic the anti-Arrhenius temperature dependence of the folding rates seen in proteins[114]. 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

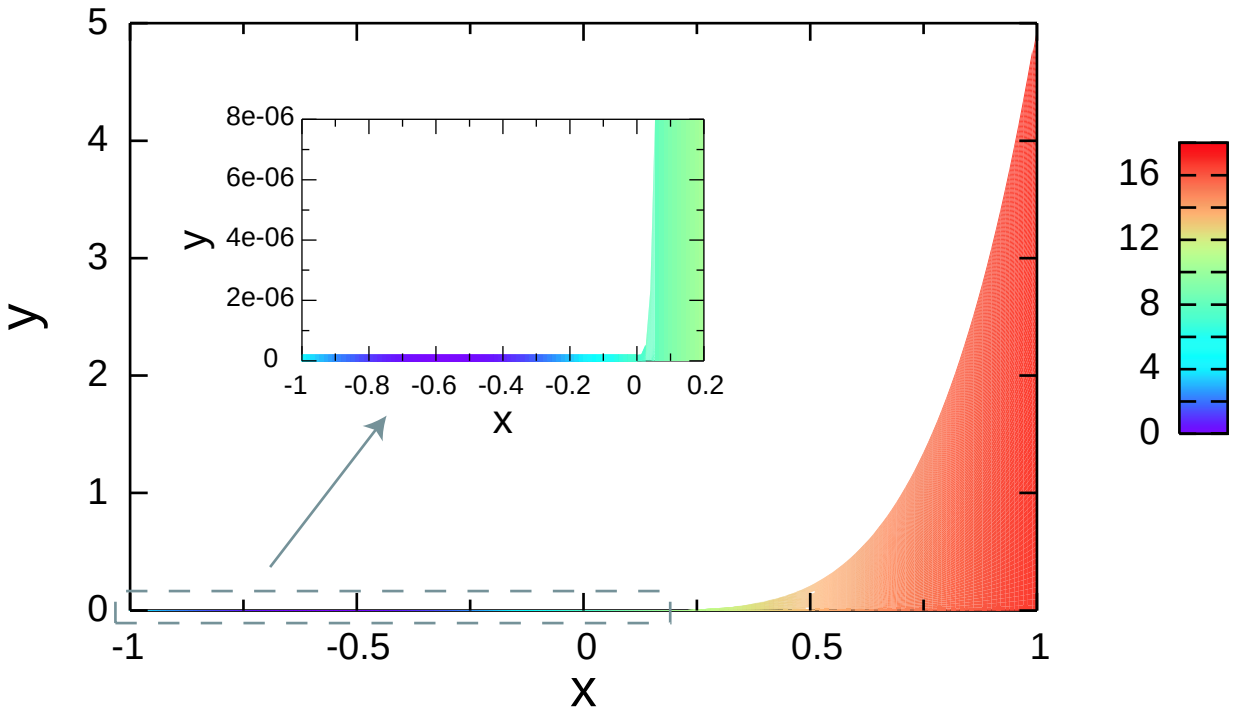

Figure 4.1: A schematic representation of the two-dimensional potential function used in previous chapter. The colored area corresponds to the accessible region of the (x, y) plane, with the colors representing the magnitude of the potential energy at that (x, y) point (scale bar in kcal/mol). The potential energy is infinite in the non-colored region and for $y < 0$, $x < -1$, and $x > 1$. The inset is an enlarged view of the folded macrostate and transition region.

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 $x > 0$

$$B(x) = \begin{cases} \delta & 1 \leq x \leq 0 \\ bx^{n_1} + \delta & 0 < x \leq 1 \end{cases} \tag{4.1}$$

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

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

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 $x_0 = 0.5745$, $x_1 = 0.05222$, $x_2 = 0.03830$, and the energy offset $c_1 = 5.402$ kcal/mol were chosen so that $U(x,y)$ and its first derivative are continuous.

We use Metropolis MC sampling 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 MC proposal scheme[108, 75]. The step size in the y direction varies with $B(x)$, i.e. a proposed move $(\Delta x', \Delta y')$ is generated uniformly in the region $-\Delta < \Delta x' < \Delta$, $-B(x)\Delta < \Delta y' < B(x)\Delta$, where $\Delta = 0.01$ is a constant for all temperatures. To correct for the asymmetric MC proposal 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" $-0.1 < x < 0.0437$ was defined as not belonging to either the folded or unfolded macrostate to reduce artefactual rapid recrossings of the barrier. As discussed previously[114], the folding rate has "anti-Arrhenius" behavior, i.e. it decreases as temperature increases, as shown in Fig. 4.2. Our goal is to reproduce this temperature dependence of the folding and unfolding rate using a kinetic network model.

If the system moves diffusively along a reaction coordinate x, the Fokker-Plan equation can be used to describe this stochastic motion superimposed with deterministic drift[115],

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial}{\partial x}[v(x)P - \frac{\partial}{\partial x}D(x)P]$$

where $P(x,t)$ is the probability density function of the system, $v(x)$ is the drift velocity, $D(x)$ is the diffusion coefficient. The drift and diffusion coefficient can be fully reconstructed from short-time simulation, and in turn, if a network is imposed with the same drift and diffusion coefficient along the reaction coordinate, it should return the same kinetics as that of the system.

In order to reproduce the kinetic characteristics of the 2-D system with the discrete network model we make use of the local drift velocity and diffusion coefficients. Multiple short-time MC trajectories were run at different starting points along the reaction coordinate x; the drift velocity $v(x_0)$ and diffusion coefficient $D(x_0)$ were evaluated using[51]

$$v(x_0) = \frac{\partial \langle x(t,x_0)\rangle}{\partial t}$$

and

$$D(x_0) = \frac{1}{2}\frac{\partial \sigma^2(t,x_0)}{\partial t}.$$

In practice, the derivatives are computed by fitting a straight line to $\langle x(t,x_0)\rangle$ and $\sigma^2(t,x_0)$ as a function of t. Our goal is to build up a network with the same local drift velocity and

diffusion coefficients as the MC simulation of the system, with the expectation that such a network will reproduce the kinetics of the system.

4.2.2 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 of the two-dimensional potential with $S = 8$ replicas at temperatures ranging from 296 K 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 $N = 50,000$ configurations at each temperature, and $N \times S = 400,000$ 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 not too far from the simulated temperatures.

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 have shown[75] 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 re-weighting scheme designed to minimize statistical error.[75] The T-WHAM canonical average $\langle A(T_0)\rangle$ of a quantity A at temperature T_0 is

$$\langle A(T_0)\rangle = \sum_{i=1}^{N} w_i(T_0) A_i, \tag{4.2}$$

where the summation runs over the N RE conformations from all temperatures, A_i is the

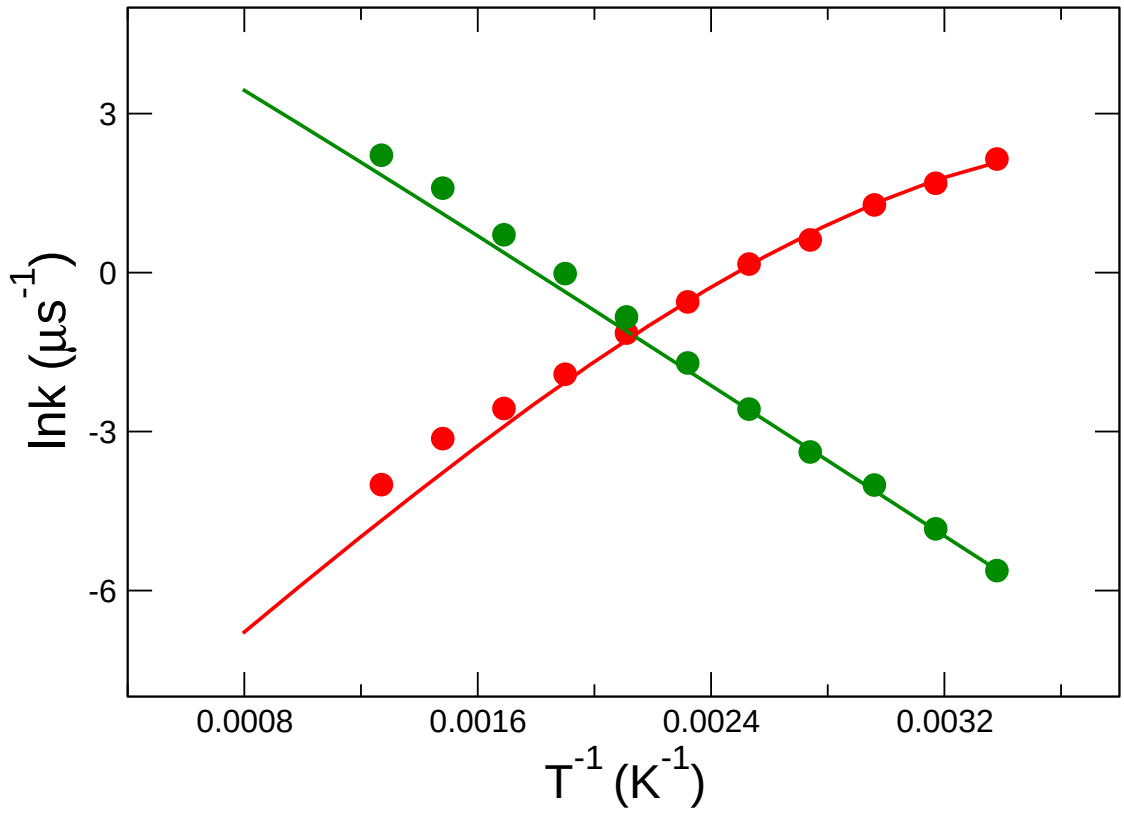

Figure 4.2: The temperature dependence of the folding and unfolding rate constants. Folding and unfolding rates are indicated by red and green, respectively. The rate constants indicated by circles were derived from kinetic MC simulation run at different temperatures. The lines represent the rates calculated 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.

value of A for conformation i, and the weight factor $w_i(T_0)$ is given by

$$w_i(T_0) = \left\{ \sum_{k=1}^{S} N_k f_k \exp\left[\left(\frac{1}{k_b T_0} - \frac{1}{k_b T_k} \right) E_i \right] \right\}^{-1}, \tag{4.3}$$

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 constants f_k in Eq. 4.3 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 non-linear equations known as the WHAM equations [75, 116]. Thus, each sample i has a weight factor associated with it (Eq. 4.3) that depends only on its energy E_i and the temperature of interest T_0, and *not* at the temperature the sample was originally collected. To calculate the PMF of the system as a function of x at temperature T_0 using the discretized state space, it is sufficient to employ Eq. 4.2 with A being an indicator function which is non-zero if the x-coordinate of the sample is near the designated value of x. This can be done for any temperature T_0, which needs not be one of the temperatures used in the RE simulation. In Fig. 4.3, the potential of mean force (PMF) calculated using the weight factors matches perfectly with that evaluated directly from the function form.

4.2.3 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. 4.4 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 ergodicity and 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 cut-off 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

$$\frac{k_{ij}(T)}{k_{ji}(T)} = \frac{w_i(T)}{w_j(T)}, \tag{4.4}$$

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, $k_{ji}(T)$ 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[117]. During the simulation, the population histogram along the x coordinate (the reaction coordinate for our two-dimensional system) was 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.

4.2.4 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

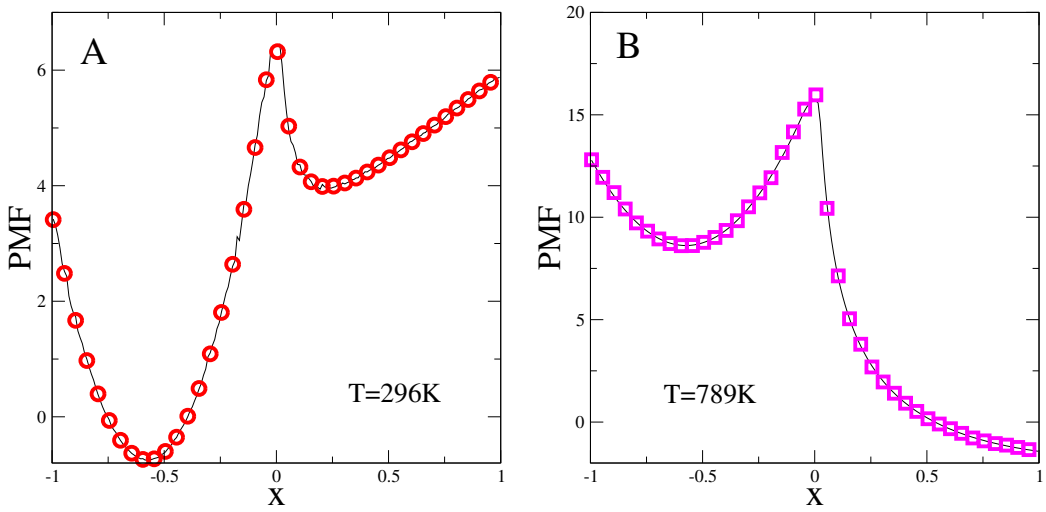

Figure 4.3: The PMF at two different temperature 296 K and 789 K. Solid lines are the exact value calculated by numerical integration of the potential. Circles are derived from the full ensemble of 8 temperatures combined using WHAM.

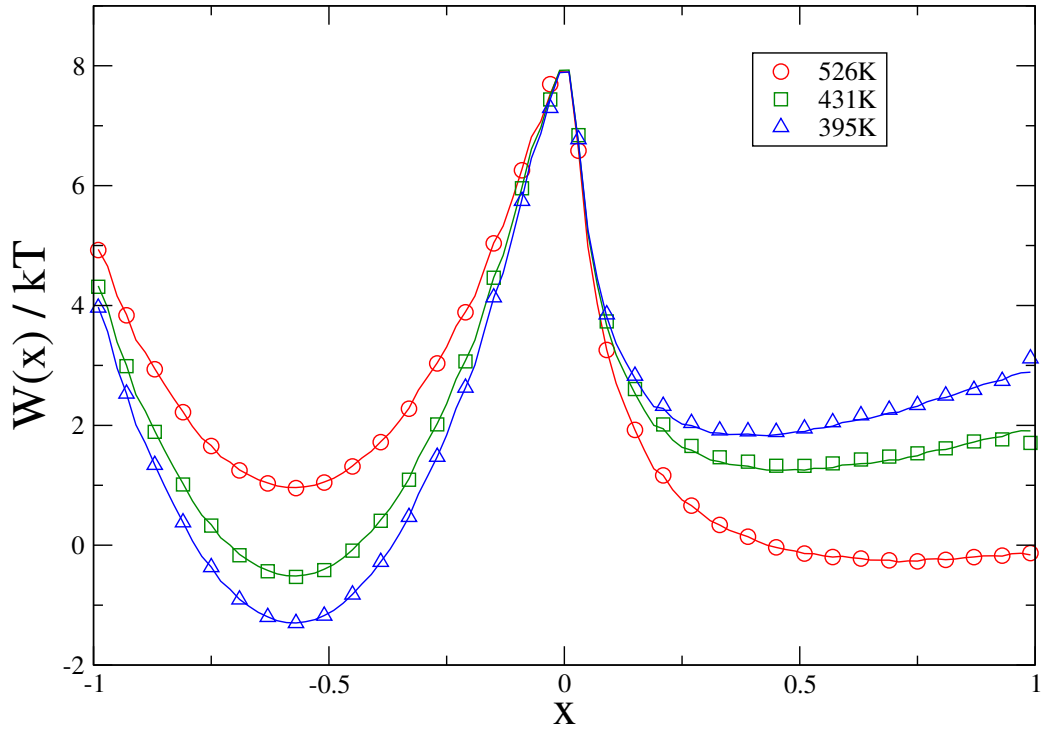

Figure 4.4: The PMF along the x coordinate at the three temperatures 395 K, 431 K, and 526 K (blue, green, and red, respectively). Solid lines are the exact PMFs calculated by numerical integration of the potential, while the circles are derived from kinetic network simulations at each temperature.

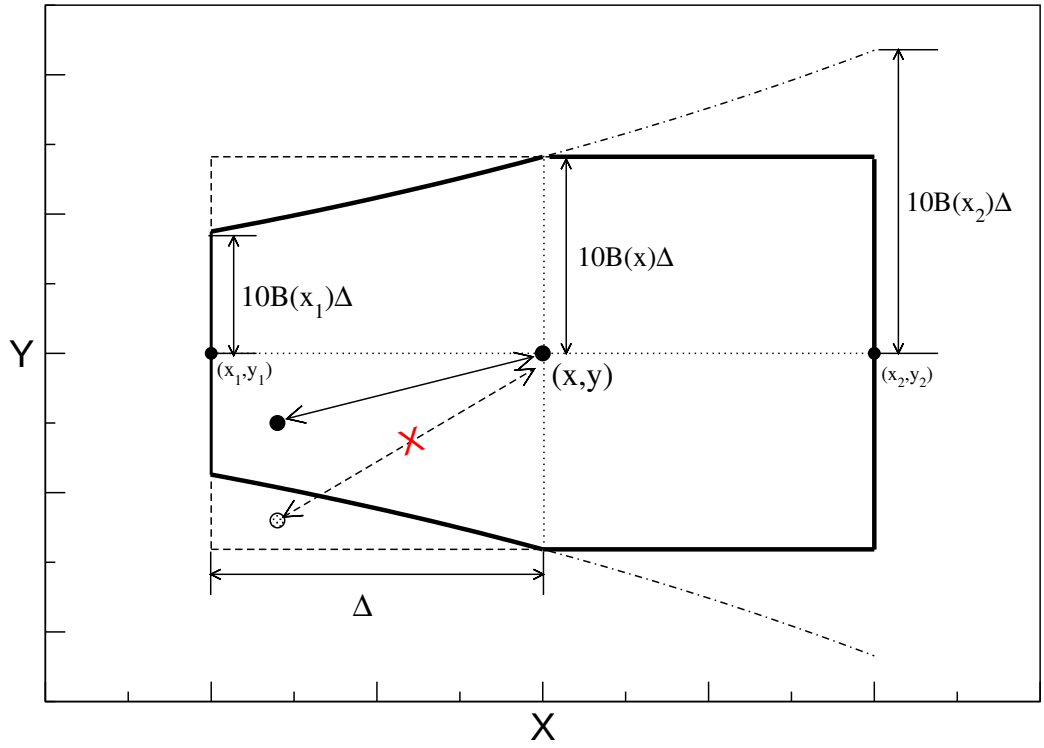

Figure 4.5: Diagram illustrating the neighboring pair rule for the network model, showing the locus of points (region within the solid line) that can be reversibly visited from a given reference point (x, y). $B(x)$ is the function that defines the accessible region of the system, Δ is the maximum MC step size, x_1, y_1, x_2, 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,

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

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

and

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

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. 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 MC, we ran 10,000 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 Fig. 4.6.

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 an 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. If all of these outgoing rates are similar, then the waiting time in a given node will be approximately proportional to the inverse of the number of neighbors of that node. As seen in Fig. 4.7, the average number of neighbors for a node increases with x due to the bigger cut-off length in y direction used to define network edges. Thus, the average waiting time between transitions among microstates will 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} is specific to the use of kinetic MC as a reference dynamical simulation method on the continuous potential, and will likely not generalize to Newtonian dynamics on a high-dimensional potential.

4.3 Results and Discussion

To confirm that the $400,000$ 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. 4.3 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 Fig. 3.3 for clarity). This indicates that the correctly weighted discretized state space is a

good approximation to the PMF on the continuous potential at all 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 Fig. 4.4. 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 the conditional potential, we obtained optimized rate parameters for the network model. We found that the choice of μ_{ij} described above with $c_0 = 0.85$ at $T_0 = 298$ K (for all x) gives good agreement, as shown in Fig. 4.6. Furthermore, the folding rates at different temperature obtained from MC simulations on the continuous potential and from the discretized kinetic network simulation agree very well, as shown in Fig. 4.8.

We have previously shown 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.[74] to model peptide folding using an all-atom potential function with hundreds of degrees of freedom, while retaining the correct thermodynamic behavior. The network model of Andrec, et al. 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.[75] In fact, the present formulation yields correct PMFs with respect to any choice of reduced coordinate. This is because the f_k factors which appear in Eq. 4.3 are free energies associated with a given replica, and are in principle independent of the choice of reduced coordinate. While the WHAM equations themselves require a choice of reduced coordinate which one uses to construct the histograms, the resulting f_k factors do

not depend on that choice. While 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 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[49, 50, 51, 52, 53, 54] 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[74] can yield physically plausible global kinetics of molecular systems. Taken together, these observations indicate that the methodology described here will 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 drift and diffusion parameters could be obtained. In general, this can lead to local dynamics which is heterogeneous. This issue could be overcome by the partitioning of nodes into clusters, which could be done based on local dynamical parameters, or more simply, on structural considerations. Drift and diffusion parameters could then be estimated separately for each cluster along the reduced coordinate, accounting for the heterogeneity without the need for approximating kinetics in a multidimensional space. Furthermore, the drift and diffusion can be calculated using generalized coordinates, or the calibration of the network model parameters could 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 which they connect) was straightforwardly dictated by the move set of the kinetic scheme MC 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 will require a multidimensional search over topology and rate parameters μ_{ij}.

4.4 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[74]. Our network model can be viewed as combining the advantages of other methods for the study of slow kinetics, while providing mechanisms for avoiding some of their pitfalls. As in previous methods[49, 50, 51, 52, 53, 54], we compute local stochastic dynamical quantities on a one-dimensional reaction coordinate, but only as a benchmark to calibrate the rate parameters of a network model constructed from the full discretized state space of the system. However, 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. The kinetic simulations are performed not on a reduced low-dimensional landscape, but on a network that can allow for multiple reaction pathways. This gives us the flexibility to visualize the dynamics on reaction coordinates of our choosing. The network model is a Markovian model, like that of other previous approaches[60, 61, 62, 63, 65], but instead of using artificially defined macrostates, we use a large number of microstates collected from an 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. Nonetheless, 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

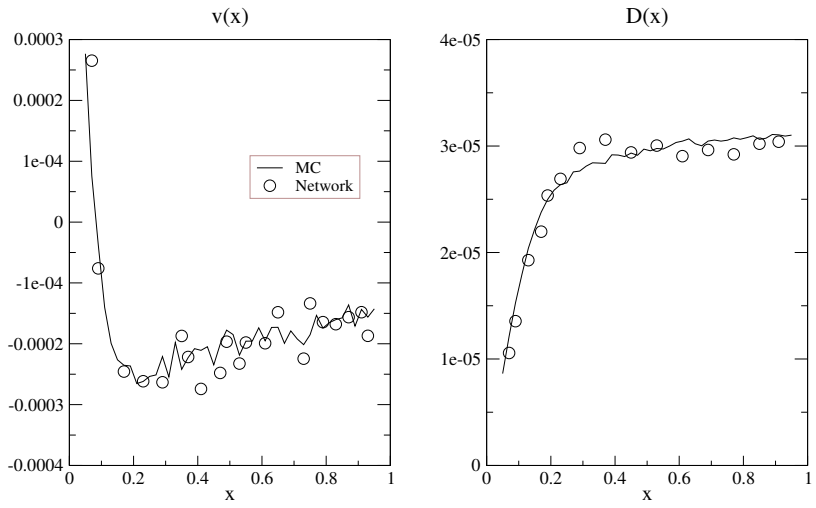

Figure 4.6: The drift velocity $v(x)$ and diffusion coefficient $D(x)$ along the reaction coordinate x at 298 K. The lines represent the drift velocity and diffusion coefficient of the kinetic MC simulation, while the circles are the results from the kinetic network model after calibration of c_0.

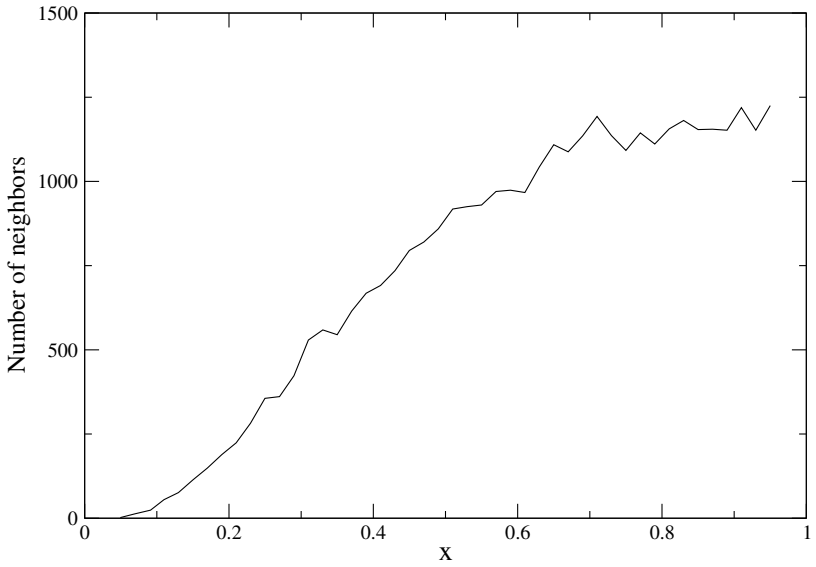

Figure 4.7: The average number of neighbors per node for all nodes which have a given value of the reaction coordinate x.

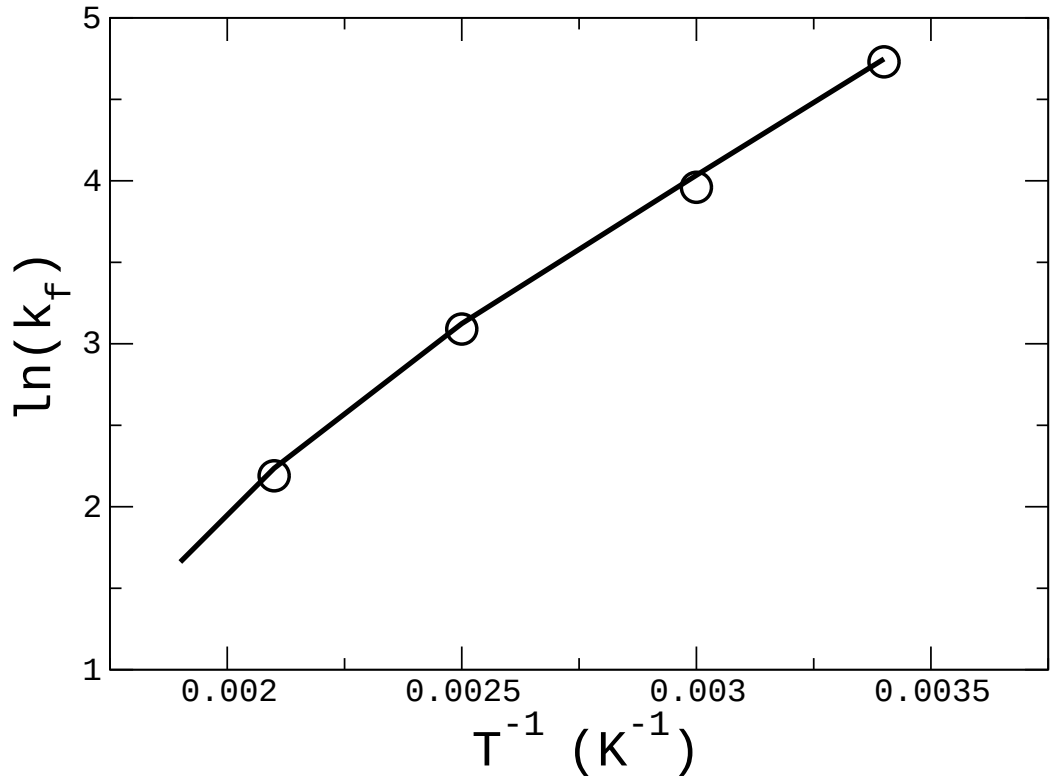

Figure 4.8: Arrhenius plot of the folding rates of the model system. The line represents the folding rate from kinetic MC simulation in unit of 10^{-6} per MC step. The circles represent the rates from simulation of the kinetic network model.

forces do not need to be evaluated. If necessary, methods for accelerating Gillespie-type simulations that have been developed in the context of chemical reaction and systems biology simulations could be used to mitigate the computational burden[118]. 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.

4.5 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, i.e. which pairs of nodes are to be connected by edges. In previous work[74], 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 non-reversibility of the particular move set used in the MC simulation. In Fig. 4.5 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

$$\begin{aligned} x - \Delta < x' < x \qquad &\text{and} \qquad |y - y'| < 10B(x')\Delta \\ x < x' < x + \Delta \qquad &\text{and} \qquad |y' - y| < 10B(x)\Delta\,. \end{aligned} \tag{4.5}$$

References

[1] J.N. Onuchic and P.G. Wolynes. Theory of protein folding. *Curr. Opin. Struct. Biol.*, 14:70–75, 2004.

[2] Yuko Okamoto. Generalized-ensemble algorithms: enhanced sampling techniques for monte carlo and molecular dynamics simulations. *Journal of Molecular Graphics and Modelling*, 22:425–439, 2004.

[3] G. Torrie and J. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *J. Comp. Phys.*, 23:187, 1977.

[4] A. M. Ferrenberg and R. H. Swendsen. Optimized monte carlo data analysis. *Phys. Rev. Lett.*, 63:1195, 1989.

[5] Joan-Emma Shea and Charles L. Brooks III. From folding theories to folding proteins: A review and assessment of simulation studies of protein folding and unfolding. *Annu. Rev. Phys. Chem.*, 52:499–535, 2001.

[6] Krishna Pratap Ravindranathan, Emilio Gallicchio, and Ronald M. Levy. Conformational equilibria and free energy profiles for the allosteric transition of the ribose-binding protein. *J. Mol. Biol.*, 353:196–210, 2005.

[7] Bernd A. Berg and Thomas Neuhaus. Multicanonical algorithms for first order phase transitions. *Phys. Lett. B*, 267:249–253, 1991.

[8] Ulrich H. E. Hansmann and Yuko Okamoto. Prediction of peptide conformation by multicanonical algorithm: New approach to the multiple-minimum problem. *J. Comp. Chem.*, 14:1333–1338, 1993.

[9] C. Bartels and M. Karplus. Probability distributions for complex systems: Adaptive umbrella sampling of the potential energy. *J. Phys. Chem. B.*, 102:865, 1998.

[10] E. Marinari and G. Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.*, 19:451–458, 1992.

[11] Robert H. Swendsen and Jian-Sheng Wang. Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.*, 57:2607–2609, 1986.

[12] Koji Hukushima and Koji Nemoto. Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. Jpn.*, 65:1604–1608, 1996.

[13] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculation by fast computing machines. *J. Chem. Phys.*, 21:1087–1091, 1953.

[14] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314:141–151, 1999.

[15] Young Min Rhee and Vijay S. Pande. Multiplexed-replica exchange molecular dynamics method for protein folding simulations. *Biophys. J.*, 84:775–786, 2003.

[16] Hugh Nymeyer, S. Gnanakaran, and Angel E. García. Atomic simulations of protein folding, using the replica exchange algorithm. *Meth. Enzymol.*, 383:119–149, 2004.

[17] M. Cecchini, F. Rao, M. Seeber, and A. Caflisch. Replica exchange molecular dynamics simulations of amyloid peptide aggregation. *J. Chem. Phys.*, 121:10748, 2004.

[18] Hui-Hsu (Gavin) Tsai, Meital Reches, Chung-Jung Tsai, Kannan Gunasekaran, Ehud Gazit, and Ruth Nussinov. Energy landscape of amyloidogenic peptide oligomerization by parallel-tempering molecular dynamics simulation: Significant role of Asn ladder. *Proc. Natl. Acad. Sci. USA*, 102:8174–8179, 2005.

[19] A. Baumketner and J.-E. Shea. Free energy landscapes for amylodogenic tetrapeptides dimerization. *Biophys. J.*, 89:1493–1503, 2005.

[20] Gennady M. Verkhivker, Paulo A. Rejto, Djamal Bouzida, Sandra Arthurs, Anthony B. Colson, Stephan T. Freer, Daniel K. Gehlhaar, Veda Larson, Brock A. Luty, Tami Marrone, and Peter W. Rose. Parallel simulated tempering dynamics of ligand-protein binding with ensembles of protein conformations. *Chem. Phys. Lett.*, 337:181–189, 2001.

[21] Krishna Pratap Ravindranathan, Emilio Gallicchio, Richard A. Friesner, Ann E. McDermott, and Ronald M. Levy. Conformational equilibrium of cytochrome P450 BM-3 complexed with N-palmitoylglycine: A replica exchange molecular dynamics study. *J. Am. Chem. Soc.*, 128:5786–5791, 2006.

[22] Francesco Rao and Amadeo Caflisch. Replica exchange molecular dynamics simulations of reversible folding. *J. Chem. Phys.*, 119:4035–4042, 2003.

[23] M. Marvin Seibert, Alexandra Patriksson, Berk Hess, and David van der Spoel. Reproducible polypeptide folding and structure prediction using molecular dynamics simulations. *J. Mol. Biol.*, 354:173–183, 2005.

[24] David A. Kofke. On the acceptance probability of replica-exchange Monte Carlo trials. *J. Chem. Phys.*, 117:6911–6914, 2002.

[25] Aminata Kone and David A. Kofke. Selection of temperature intervals for parallel-tempering simulations. *J. Chem. Phys.*, 122:206101, 2005.

[26] Cristian Predescu, Mihaela Predescu, and Cristian V. Ciobanu. The incomplete beta function law for parallel tempering sampling of classical canonical systems. *J. Chem. Phys.*, 120:4119–4128, 2004.

[27] Cristian Predescu, Mihaela Predescu, and Cristian V. Ciobanu. On the efficiency of exchange in parallel tempering Monte Carlo simulations. *J. Phys. Chem. B*, 109:4189–4196, 2005.

[28] Nitin Rathore, Manan Chopra, and Juan J. de Pablo. Optimal allocation of replicas in parallel tempering simulations. *J. Chem. Phys.*, 122:024111, 2005.

[29] Simon Trebst, Matthias Troyer, and Ulrich H. E. Hansmann. Optimized parallel tempering simulations of proteins. *J. Chem. Phys.*, 124:174903, 2006.

[30] Daniel M. Zuckerman and Edward Lyman. A second look at canonical sampling of biomolecules using replica exchange simulation. *J. Chem. Theory Comput.*, 2:1200–1202, 2006.

[31] Daniel M. Zuckerman. Erratum to "A second look at canonical sampling of biomolecules uisng replica exchange simulation". 2006.

[32] David A. C. Beck, George W. N. White, and Valerie Daggett. Exploring the energy landscape of protein folding using replica-exchange and conventional molecular dynamics simulations. *J. Struct. Biol.*, 157:514–523, 2007.

[33] Shin-Ichi Segawa and Mitsuru Sugihara. Characterization of the transition state of lysozyme unfolding. I. Effect of protein-solvent interactions on the transition state. *Biopolymers*, 23:2473–2488, 1984.

[34] Mikael Oliveberg, Yee-Joo Tan, and Alan R. Fersht. Negative activation enthalpies in the kinetics of protein folding. *Proc. Natl. Acad. Sci. USA*, 92:8926–8929, 1995.

[35] V. Muñoz, P. A. Thompson, J. Hofrichter, and W. A. Eaton. Folding dynamics and mechanism of β-hairpin formation. *Nature*, 390:196–199, 1997.

[36] Martin Karplus. Aspects of protein reaction dynamics: Deviations from simple behavior. *J. Phys. Chem. B*, 104:11–27, 2000.

[37] Philippe Ferrara, Joannis Apostolakis, and Amadeo Caflisch. Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations. *J. Phys. Chem. B*, 104:5000–5010, 2000.

[38] Wei Yuan Yang and Martin Gruebele. Rate—temperature relationships in λ-repressor fragment λ_{6-85} folding. *Biochemistry*, 43:13018–13025, 2004.

[39] Scalley ML and Baker D. Protein folding kinetics exhibit an arrhenius temperature dependence when corrected for the temperature dependence of protein stability. *Proc. Natl. Acad. Sci. USA*, 94:10636–10640, 1997.

[40] J.D . Bryngelson and P.G. Wolynes. Intermediates and barrier crossing in a randomenergymodel (with applications to protein folding). *J.Phys.Chem.*, 93:6902, 1989.

[41] Ron Elber. Long-timescale simulation methods. *Curr. Opinion Struct. Biol.*, 15:151–156, 2005.

[42] Christopher D. Snow, Eric J. Sorin, Young Min Rhee, and Vijay S. Pande. How well can simulation predict protein folding kinetics and thermodynamics? *Ann. Rev. Biophys. Biomol. Struct.*, 34:43–69, 2005.

[43] Markus Christen and Wilfred F. van Gunsteren. On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: A review. *J. Comput. Chem.*, 29:157–166, 2007.

[44] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.*, 53:291–318, 2002.

[45] B. Zagrovic, E. J. Sorin, and V. Pande. β-hairpin folding simulations in atomistic detail using an implicit solvent model. *J. Mol. Biol.*, 313:151–169, 2001.

[46] V. S. Pande, I. Baker, J. Chapman, S. P. Elmer, S. Khaliq, S. M. Larson, Y. M. Rhee, M. R. Shirts, C. D. Snow, E. J. Sorin, and B. Zagrovic. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers*, 68:91–109, 2003.

[47] Anton K. Faradjian and Ron Elber. Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.*, 120:10880, 2004.

[48] A. R. Fersht. On the simulation of protein folding by short time scale molecular dynamics and distributed computing. *Proc. Natl. Acad. Sci. USA*, 99:14122–14125, 2002.

[49] Mark F. Schumaker, Régis Pomès, and Benoît Roux. A combined molecular dynamics and diffusion model of single proton conduction through gramacidin. *Biophys. J.*, 79:2840–2857, 2000.

[50] Gerhard Hummer and Ioannis G. Kevrekidis. Coarse molecular dynamics of a peptide fragment: Free energy, kinetics and long-time dynamics computations. *J. Chem. Phys.*, 118:10762, 2003.

[51] Dmitry I. Kopelevich, Athanassios Z. Panagiotopoulos, and Ioannis G. Kevrekidis. Coarse-grained kinetic computations of rare events: Application to micelle formation. *J. Chem. Phys.*, 122:044908, 2005.

[52] Robert B. Best and Gerhard Hummer. Diffusive model of protein folding dynamics with Kramers turnover in rate. *Phys. Rev. Lett.*, 96:228104, 2006.

[53] Sichun Yang, Josè N. Onuchic, and Herbert Levine. Effective stochastic dynamics on a protein folding energy landscape. *J. Chem. Phys.*, 125:054910, 2006.

[54] Sichun Yang, Josè N. Onuchic, Angel E. Garcìa, and Herbert Levine. Folding time predictions from all-atom replica exchange simulations. *J. Mol. Biol.*, 372:756–763, 2007.

[55] Ao Ma and Aaron R. Dinner. Automatic method for identifying reaction coordinates in complex systems. *J. Phys. Chem. B*, 109:6769–6779, 2005.

[56] Sergei V. Krivov and Martin Karplus. One-dimensional free-energy profiles of complex systems: Progress variables that preserve barriers. *J. Phys. Chem. B*, 110:12689–12698, 2006.

[57] Sergei V. Krivov, Stefanie Muff, Amadeo Caflisch, and Martin Karplus. One-dimensional barrier-preserving free-energy projections of a β-sheet miniprotein: New insights into the folding process. *J. Phys. Chem. B*, 112:8701–8714, 2008.

[58] F. Rao and A. Caflisch. The protein folding network. *J. Mol. Biol.*, 342:299–306, 2004.

[59] S. B. Ozkan, K. A. Dill, and I. Bahar. Fast-folding protein kinetics, hidden intermediates, and the sequential stabilization model. *Protein Sci.*, 11:1958–1970, 2002.

[60] Y.-J. Ye, D. R. Ripoll, and H. A. Scheraga. Kinetics of cooperative protein folding involving two separate conformational families. *Comp. Theor. Polymer Sci.*, 9:359–370, 1999.

[61] N. Singhal, C. D. Snow, and V. S. Pande. Using path sampling to build better Markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chem. Phys.*, 121:415–425, 2004.

[62] W. C. Swope, J. W. Pitera, and F. Suits. Describing protein folding kinetics by molecular dynamics simulations. 1. Theory. *J. Phys. Chem. B*, 108:6571–6581, 2004.

[63] W. C. Swope, J. W. Pitera, F. Suits, M. Pitman, M. Eleftheriou, B. G. Fitch, R. S. Germain, A. Rayshubski, T. J. C. Ward, Y. Zhestkov, and R. Zhou. Describing protein folding kinetics by molecular dynamics simulations. 2. Example applications to alanine dipeptide and a β-hairpin peptide. *J. Phys. Chem. B*, 108:6582–6594, 2004.

[64] D. S. Chekmarev, T. Ishida, and R. M. Levy. Long time conformational transitions of alanine dipeptide in aqueous solution: Continuous and discrete state kinetic models. *J. Phys. Chem.*, 108:19487–19495, 2004.

[65] Nicolae-Viorel Buchete and Gerhard Hummer. Peptide folding kinetics from replica exchange molecular dynamics. *Phys. Rev. E*, 77:030902, 2008.

[66] G. Wei, N. Mousseau, and P. Derreumaux. Complex folding pathways in a simple β-hairpin. *Proteins*, 56:464–474, 2004.

[67] D. A. Evans and D. J. Wales. Folding of the GB1 hairpin peptide from discrete path sampling. *J. Chem. Phys.*, 121:1080–1090, 2004.

[68] David J. Wales. Discrete path sampling. *Mol. Phys.*, 100:3285–3305, 2002.

[69] Joanne M. Carr and David J. Wales. Folding pathways and rates for the three-stranded β-sheet peptide Beta3s using discrete path sampling. *J. Phys. Chem. B*, 112:8760–8769, 2008.

[70] Sergei V. Krivov and Martin Karplus. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. USA*, 101(41):14766–14770, 2004.

[71] A. Mitsutake, Y. Sugita, and Y. Okamoto. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers*, 60:96–123, 2001.

[72] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314:141–151, 1999.

[73] D. van der Spoel and M.M. Seibert. Protein folding kinetics and thermodynamics from atomistic simulations. *Phys.Rev.Lett.*, 96:238102, 2006.

[74] M. Andrec, A. K. Felts, E. Gallicchio, and R. M. Levy. Protein folding pathways from replica exchange simulations and a kinetic network model. *Proc. Natl. Acad. Sci. USA*, 102:6801–6806, 2005.

[75] E. Gallicchio, M. Andrec, A. K. Felts, and R. M. Levy. Temperature weighted histogram analysis method, replica exchange and transition paths. *J. Phys. Chem. B*, 109:6722–6731, 2005.

[76] F.Rao and A.Caflisch. The protein folding network. *J. Mol. Biol.*, 342:299–306, 2004.

[77] K.A.Dill S.B.Ozkan and I.Bahar. Fast-folding protein kinetics, hidden intermediates, and the sequential stabilization model. *Protein Sci.*, 11:1958–1970, 2002.

[78] K.A.Dill S.B.Ozkan and I.Bahar. Computing the transition state populations in simple protein models. *Biopolymers*, 68:35–46, 2003.

[79] W. C . Swope, J. W. Pitera, and F. Suits. Describing protein folding kinetics by molecular dynamics simulations. 1. theory. *J.Phys.Chem.B*, 108:6571–6581, 2004.

[80] W. C . Swope, J. W. Pitera, and F. etal. Suits. Describing protein folding kinetics by molecular dynamics simulations. 2. example applications to alanine dipeptide and a -hairpin peptide. *J.Phys.Chem.B*, 108:6582–6594, 2004.

[81] Nina Singhal, Christopher D. Snow, and Vijay Pande. Using path sampling to build better Markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J.Chem.Phys.*, 121:415–425, 2004.

[82] Y.-J . Ye, D.R. Ripoll, and H.A. Scheraga. Kinetics of cooperative protein folding involving two separate conformational families. *Comp.Theor.Polymer Sci.*, 9:359–370, 1999.

[83] D.S . Chekmarev, T. Ishida, and R.M. Levy. Long-time conformational transitions of alanine dipeptide in aqueous solution: Continuous and discrete-state kinetic models. *J.Phys.Chem.B*, 108:19487–19495, 2004.

[84] G . Wei, N. Mousseau, and P. Derreumaux. Complex folding pathways in a simple beta-hairpin. *Proteins*, 56:464–474, 2004.

[85] D.A. Evans and D.J. Wales. Folding of the gb1 hairpin peptide from discrete path sampling. *J.Chem.Phys.*, 121:1080–1090, 2004.

[86] D.T.Gillespie. *Markov Process: An Introduction to Physicall Scientists*, 1992.

[87] D.A. McQuarrie and JD Simon. *Physical Chemistry: A Molecular Approach*, 1997.

[88] Morten Hagen, Byungchan Kim, Pu Liu, Richard A. Friesner, and B. J. Berne. Serial replica exchange. *J. Phys. Chem. B*, 111:1416–1423, 2007.

[89] E . Gallicchio, M . Andrec, A.K. Felts, and R.M. Levy. Temperature weighted histogram analysis method, replica exchange, and transition paths. *J.Phys.Chem.B*, 109:6722–6731, 2005.

[90] Pu Liu, Byungchan Kim, Richard A. Friesner, and B. J. Berne. Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proc. Natl. Acad. Sci. USA*, 102:13749–13754, 2005.

[91] Charles J. Geyer and Elizabeth A. Thompson. Annealing Markov chain Monte Carlo with application to ancestral inference. *J. Am. Stat. Assoc.*, 90:909–920, 1995.

[92] Ulrich H. E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, 281:140–150, 1997.

[93] Yuji Sugita, Akio Kitao, and Yuko Okamoto. Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.*, 113:6042–6051, 2000.

[94] H . Fukunishi, O. Watanabe, and S. Takada. *J.Chem.Phys*, 116:9058–9067, 2002.

[95] W . Kwak and U.H.E. Hansmann. Efficient sampling of protein structures by model hopping. *Phys.Rev.Lett.*, 95:138102, 2005.

[96] Liu P ., X . Huang, R. Zhou, and B.J. Berne. Hydrophobic aided replica exchange: an efficient algorithm for protein folding in explicit solvent. *J.Phys.Chem.B*, 110:19018–19022, 2006.

[97] D . Min, H . Li, G . Li, R. Bitetti-Putzer, and W. Yang. Dual-topology hamiltonian-replica-exchange overlap histogramming method to calculate relative free energy difference in rough energy landscape. *arXiv:physics 0605005*, 2006.

[98] Weihua Zheng, Michael Andrec, Emilio Gallicchio, and Ronald M. Levy. Simulating Replica Exchange simulations of protein folding using a kinetic network model. *Proc. Natl. Acad. Sci. USA*, 104:15340–15345, 2007.

[99] W. Nadler and U.H.E. Hansmann. On dynamics and optimal number of replicas in paralle tempering simulations. *arXiv:0709.3289v1*, 2008.

[100] Michael Andrec, Anthony K. Felts, Emilio Gallicchio, and Ronald M. Levy. Protein folding pathways from replica exchange simulations and a kinetic network model. *Proc. Natl. Acad. Sci. USA*, 102:6801–6806, 2005.

[101] W . Zhang, C. Wu, and Y. Duan. Convergence of replica exchange molecular dynamics. *J.Chem.Phys.*, 123:154105, 2005.

[102] X. Periole and A.E. Mark. Convergence and sampling efficiency in replica exchange simulations of peptide folding in explicit solvent. *J.Chem.Phys.*, 126:014903, 2007.

[103] B . Bicout and A. Szabo. Entropic barriers, transition states, funnels, and exponential protein folding kinetics: a simple model. *Protein Sci.*, 9:452–465, 2000.

[104] H.-X Zhou and R. Zwanzig. A rate process with an entropy barrier. *J.Chem.Phys.*, 94:6147, 1991.

[105] H.S . Chan and K.A. Dill. Protein folding in the landscape perspective:chevron plots and non-arrhenius kinetics. *Proteins*, 30:2, 1998.

[106] A . Kolinski and J. Skolnick. *Polymer*, 45:511, 2004.

[107] G. Tiana, L.Sutto, and R.A. Broglia. Use of the Metropolis algorithm to simulate the dynamics of protein chains. *Physica A*, 380:241–249, 2007.

[108] M.P . Allen and D.J. Tildesley. *Computer Simulation of Liquids*, Oxford, 1987.

[109] D. Chandler. Statistical mechanics of isomerization dynamics in liquids and the transition state approximation. *J.Chem.Phys.*, 68:2959–2970, 1978.

[110] R.M . Levy, M . Karplus, and J.A. McCammon. Diffusive langevin dynamics of model alkanes. *Chem.Phys.Lett.*, 65:4–11, 1979.

[111] J . Borreguero, N . Dokholyan, S . Buldyrev, E . Shakhnovich, and H. Stanley. Thermodynamics and folding kinetics analysis of the sh3 domain form discrete molecular dynamics. *J.Mol.Biol.*, 318:863–876, 2002.

[112] S . Park and V.S. Pande. Validation of markov state models using shannon's entropy. *J.Chem.Phys.*, 124:054118, 2006.

[113] F . Noé, I . Horenko, C . Schütte, and J.C. Smith. *J.Chem.Phys.*, 126:155102, 2007.

[114] Weihua Zheng, Michael Andrec, Emilio Gallicchio, and Ronald M. Levy. Simple continuous and discrete models for simulating replica exchange simulations of protein folding. *J. Phys. Chem. B*, 112:6083–6093, 2008.

[115] C.W.Gardiner. *Handbook of Stochastic Methods*, pages 3rd ed. (Springer, New York, 2004).

[116] Shankar Kumar, Djamal Bouzida, Robert H. Swendsen, Peter A. Kollman, and John M. Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *J. Comp. Chem.*, 13:1011–1021, 1992.

[117] D. T. Gillespie. *Markov Processes: An Introduction for Physical Scientists*. Academic Press, Boston, 1992.

[118] Daniel T. Gillespie. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, 58:35–55, 2007.

Vita

Weihua Zheng

2002-2009 Ph.D. in Biophysics, Rutgers University, Piscataway, NJ, USA

1999-2002 MS in Condensed Matter Physics, Univ. of Sci.&Tech. of China, Hefei, P.R.China

1994-1999 BS in Physics, Univ. of Sci.&Tech. of China, Hefei, P.R.China

Publications

In Dr. Levy's group:

Zheng W., M. Andrec, E. Gallicchio and R. M. Levy, Simple Continuous and Discrete Models for Simulating Replica Exchange Simulations of Protein Folding, *J. Phys. Chem. B* **2008**, *112*, 6083-6093.

Zheng W., M. Andrec, E. Gallicchio and R. M. Levy, Simulating Replica Exchange Simulations of Protein Folding Using a Kinetic Network Model, *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 15340.

Previous ones:

Gao H.P., Wu B.M., Wei Y.Y., Li B. and Zheng W., Lattice effects related to metal-insulator and charge-ordered transitions in $La_{0.7}Ca_{0.3}Mn_{0.9}Cr_{0.1}O_3$ and $La_{0.7}Ca_{0.3}Mn_{0.99}Cr_{0.01}O_3$, *Phys. Lett. A* **2007**, *368*, 125.

Wu B. M., M. Ausloos, Du Y. L., Zheng, W., etc., Spin Glass Behavior and Spin-Dependent Scattering in $La_{0.7}Ca_{0.3}Mn_{0.9}Cr_{0.1}O_3$ Perovskites, *Chin. Phys. Lett.*, **2005**, *22*, 686.

Wu B. M., Li, B., Zheng, W., etc., Spin-cluster effect and lattice-deformation-induced Kondo effect, Spin-glass freezing, and strong phonon scattering in La0.7Ca0.3Mn1-xCrxO3, *J. Appl. Phys.*,**2005**, *97*, 103908.

Wu B. M., M. Ausloos, Du Y. L., Zheng, W. etc., Spin Glass Behavior and Spin-Dependent Scattering in $La_{0.7}Ca_{0.3}Mn_{0.9}Cr_{0.1}O_3$ Perovskites, *Chin. Phys. Lett.*, **2005**, *22*, 686.

Li S. Y., Mo W. Q., Zheng, W., etc., Thermopower and thermal conductivity of superconducting perovskite $MgCNi_3$, *Phys. Rev. B*, **2002**, *65*, 064534.

Zheng, W., Wu B. M., Li B., Yang D. S. and Cao L. Z., Thermal conductivity in $La_{0.7}Ca_{0.3}Mn_{1-x}Cr_xO_3$ at low Temperature, *Chin. J. Low. Temp. Phys*, **2002**, *24*, 230.

Li B., Yang D. S., Zheng, W., Wu B. M. and Jin H., Electronic Transport Properties in Ce Doped Bi-2212 at Low Temperature, *Chin. J. Low. Temp. Phys.*, **2002**, *24*, 193.

Zheng, W., Wu B. M., Yang D. S., Li J. Y., Liu X. R. and Li B., Thermal Conductivity and Electrical Conductivity in $La_{1-x}Ca_xMnO_3$, *Chin. J. Low. Temp. Phys.*, **2001**, *23*, 216.

Yang D. S., Wu B. M., Zheng W., Yang H., Li B. and Cao L., Thermal Conductivity of Excess-Oxygen-Doped La_2CuO_4, *Chin. J. Low. Temp. Phys.*,**2001**, *23*, 44.