Kinetic Network Study of the Diversity and Temperature Dependence of Trp-Cage Folding Pathways: Combining Transition Path Theory with Stochastic Simulations

ARTICLE *in* THE JOURNAL OF PHYSICAL CHEMISTRY B · FEBRUARY 2011

Impact Factor: 3.3 · DOI: 10.1021/jp1089596 · Source: PubMed

5 AUTHORS, INCLUDING:

Weihua Zheng
Rice University
17 PUBLICATIONS **309** CITATIONS

Kinetic Network Study of the Diversity and Temperature Dependence of Trp-Cage Folding Pathways: Combining Transition Path Theory with Stochastic Simulations

Weihua Zheng Emilio Gallicchio Nanjie Deng Michael Andrec
Ronald M. Levy*

Department of Chemistry and Chemical Biology and
BioMaPS Institute for Quantitative Biology

Rutgers, the State University of New Jersey
Piscataway, NJ 08854

*To whom correspondence should be addressed: ronlevy@lutece.rutgers.edu

Abstract

We present a new approach to study a multitude of folding pathways and different folding mechanisms for the 20-residue mini-protein Trp-Cage using the combined power of replica exchange molecular dynamics (REMD) simulations for conformational sampling, Transition Path Theory (TPT) for constructing folding pathways and stochastic simulations for sampling the pathways in a high dimensional structure space. REMD simulations of Trp-Cage with 16 replicas at temperatures between 270K and 566K are carried out with an all-atom force field (OPLSAA) and an implicit solvent model (AGBNP). The conformations sampled from all temperatures are collected. They form a discretized state space that can be used to model the folding process. The equilibrium population for each state at a target temperature can be calculated using the Weighted-Histogram-Analysis Method (WHAM). By connecting states with similar structures and creating edges satisfying detailed balance conditions, we construct a kinetic network that preserves the equilibrium population distribution of the state space. After defining the folded and unfolded macrostates, committor probabilities (P_{fold}) are calculated by solving a set of linear equations for each node in the network and pathways are extracted together with their fluxes using the TPT algorithm. By clustering the pathways into folding "tubes", a more physically meaningful picture of the diversity of folding routes emerges. Stochastic simulations are carried out on the network and a procedure is developed to project sampled trajectories onto the folding tubes. The fluxes through the folding tubes calculated from the stochastic trajectories are in good agreement with the corresponding values obtained from the TPT analysis. The temperature dependence of the ensemble of Trp-Cage folding pathways is investigated. Above the folding temperature, a large number of diverse folding pathways with comparable fluxes flood the energy landscape. At low temperature, however, the folding transition is dominated by only a few localized pathways.

1 Introduction

Protein folding is a fundamental problem in modern molecular biophysics which occurs via rare events in a high-dimensional conformational space[1]. Computational methods can in principle provide insights into the kinetics of folding. Although reduced models[2–10] have been used to address important questions about folding kinetics, all-atom molecular simulations[11–20] are better able to explore many molecular aspects of folding. However, with current computational power, it is difficult for an all-atom MD simulation to obtain kinetics with good statistics on microsecond or longer timescales*. The kinetic details and mechanisms of how an ensemble of unfolded biomolecules with different conformations find their ways to the same native structure remain unclear.

A number of strategies for obtaining kinetic information have been proposed over the years. One approach is to sample the unfolding trajectories at high temperature. Assuming reversibility, the unfolding pathways of a stable protein should be similar to the folding pathways. Unfolding at high temperature happens in a much shorter time regime (ns). Useful but limited information can be deduced from the high temperature trajectories[12,22,23]. Another set of methods is based on the use of stochastic dynamics on a free-energy landscape[24–30]. A good reaction coordinate has to be selected to derive effective drift and diffusion coefficients along the reaction coordinate. However, for a complex process like folding, finding a small number of reaction coordinates to describe the process properly can be very difficult and may not be sufficient, key kinetic information is often lost during the projection of the system onto the chosen coordinates. A model with higher dimensionality is more likely to capture the complexity of folding pathways and kinetics. To circumvent the problems related to defining a reaction coordinate, transition path sampling (TPS) methods start with an initial transition pathway and construct dynamical pathways connecting a reactant and a product state. If an initial path is needed, extensive simulations have to be run to obtain the information. The TPS algorithm can encounter efficiency problems especially for complex systems with high barriers separating different folding routes. It is also challenging to apply TPS to large molecular systems with intervening metastable free-energy basins[31].

Another strategy for extracting kinetics consists of discretizing the state space and constructing rules for moving among those states. The resulting scheme can be represented as a graph, a roadmap or a network[32–34], the kinetics on the graph is often assumed to have Markovian behavior[35–39]. Discretization of the state space can be done by clustering based on conformational difference among structures[37,40,41], the clusters should be chosen so as to satisfy the Markovian condition[38,39,42]. Recently, transition path theory (TPT)[43,44] has been developed and was applied to model the folding of the PinWW domain[45]. A Markovian network was constructed from many relatively short MD simulations at 360K. Once the reactant and product states are defined and the committor probability (P_{fold})[46] of each state is calculated, folding pathways and their fluxes can be extracted from the network using the TPT algorithm[43,44]. The ensemble of folding pathways at 365K shows various disjoint paths leading into the native state. On the other hand, stochastic simulations on a Markovian network using the Gillespie algorithm[47] is also a powerful tool to explore the thermodynamic and kinetic properties of the network[48,49]. Numerous reactive trajectories on the network can be collected from the simulations to derive quantities such as the equilibrium population and the P_{fold} value of each node, the flux and the first passage time statistics for the reaction. While

*With the development of special purpose computers like Anton[21], microsecond molecular dynamics simulations of proteins are becoming accessible.

a single stochastic trajectory on the network is an approximation and abstraction of many all-atom trajectories in the continuous conformational space, a single pathway on the network defined in TPT theory is an abstract representation of a group of stochastic trajectories on the network. Results from stochastic simulations on the network can not only serve as a benchmark for the TPT calculation to test its validity but provide additional conformational and kinetic information.

Replica exchange molecular dynamics (REMD)[50] was developed to enhance the ability to obtain temperature canonical populations in complex systems by running many communicating simulations in parallel. The large range of temperatures of REMD enable it to achieve much better sampling at low temperatures by "borrowing" the fast kinetics at high temperatures[51]. However, since REMD involves temperature swaps between MD trajectories, it is not straightforward to obtain kinetic information from such simulations[29,39,42,52]. We have made use of a kinetic network model[53] in which we take advantage of the REMD sampling, build the nodes of the network from molecular conformations collected from REMD trajectories, then construct edges using an ansatz based on structural similarity. By allowing local transitions between two nodes that are structurally similar, we can generate trajectories or pathways that are not realized in the original REMD simulation. While this model was shown to yield physically plausible kinetics[53], the scheme we used to weight nodes arising from different simulation temperatures was such that thermodynamic parameters of the system were not exactly preserved. Recently, we presented an improved version of the kinetic network model[49] which is guaranteed to reproduce the potential of mean force (PMF) with respect to any reduced coordinates and the model was tested on a folding-like two-dimensional potential. Compared with previous work[54] which builds the Markov state model from low temperature simulations, REMD provides a more thorough search in the conformational space of the system. In this paper, we apply our network model together with both TPT[43,44] and stochastic simulations to a more complex molecular system.

The 20-residue mini-protein Trp-Cage(NLYIQ WLKDG GPSSG RPPPS), designed by Neidigh et al.[55], has been a favorite system for both computational studies and experiments[56–71]. Its native state has both a stable secondary structure and a hydrophobic core. Being a fast folder, folding events have been observed in all-atom force field molecular dynamics (MD) simulations[56,58,72,73]. REMD with different force fields and solvent models has also been used to study Trp-Cage[60,61,66,67,69,71]. Laser temperature-jump spectroscopy experiments by Qiu et al.[57] suggests that Trp-Cage is a two-state folder with folding rate $\approx (4.1\mu s)^{-1}$. Neuweiler et al.[65]showed that Trp-Cage collapses to an intermediate molten globule-like state which enhances folding efficiency. The sequence of formation between $\alpha-$helix and hydrophobic core is also of considerable interest. Both quick collapse of the core[65] and helical structure in the denatured state[64] have been observed in experiments. Juraszek et al.[66] used transition path sampling to sample dynamical pathways between folded and unfolded states of Trp-Cage and found that 80% of the paths form the tertiary contacts before helix while 20% of the paths form the helix first.

Our approach allows us to investigate a large number of folding pathways and obtain flux information of Trp-Cage in a high dimensional space with high resolution and observe the folding transitions in detail. From the results of stochastic simulations on the network, we are able to confirm the TPT formulas for the pathway fluxes. We have also studied the temperature dependent properties of the ensemble of folding paths, which has not been reported in any of the other current Markov state approaches to the best of our knowledge.

2 Methods

2.1 Constructing the kinetic network for Trp-Cage

The nodes of the kinetic network are a discretized approximation to the continuous conformational space of the system. For efficient sampling of the conformational space, we use the replica exchange molecular dynamics (REMD) method[74] with OPLS all-atom force field and the analytical generalized Born implicit solvent model (OPLSAA/AGBNP)[20,75] for simulations of Trp-Cage. The benefit of using REMD to accelerate sampling convergence at low temperature compared to low temperature molecular dynamics (MD) simulation is discussed further in the Appendix. No cutoff is set for the non-bonded interactions. A 1 fs MD time step is used. 16 replicas are applied between 270K and 566K, distributed uniformly in $1/T$. Temperature exchanges are attempted every 2ps with acceptance ratio around 40% . Each simulation starts from the NMR structure (PDB entry 1L2A) and lasts for 50ns (0.8 μs simulation time in total). Conformations are collected every 2ps from each replica for later analysis. The simulation data of the first 20ns is regarded as equilibration and excluded from further analysis. The analyzed dataset contains 240,000 conformations. This ensemble of conformations constitutes the discretized state space of Trp-Cage used in this work. The equilibrium population of each discrete state can be calculated from the T-WHAM equation as a function of temperature[76,77].

Trp-Cage has 304 atoms and 912 degrees of freedom. To reduce the number of degrees of freedom while retaining a sufficient number to describe the folding process in detail, we use a set of internal structural parameters to describe the conformations and then apply principal component analysis (PCA)[78] for dimensionality reduction. We choose a set of backbone structural parameters, 54 $C_\alpha - C_\alpha$ distances, to span the $240,000$ conformations as points in the 54-dimensional structure space. 54 internal coordinates is a lower limit for the unique determination of the relative positions of the twenty C_α atoms (60 minus 3 translational and 3 rotational degrees of freedom). The 54 distances include all possible (i, i+3), (i,i+5),(i,i+6),(i,i+14) residue pairs, plus the (2,19) and (3,18) residue pairs to account for the distance between the C-terminus and N-terminus. Principal component analysis (PCA)[78] is applied to further reduce these 54 coordinates to the twenty most significant principal components (PC), which span more than 95% of the total variance of the data. We construct the network in this twenty-dimensional PC subspace. Each PC is a linear combination of all 54 $C_\alpha - C_\alpha$ internal coordinates with the coefficients indicating the contribution of each internal coordinate to the PC. The euclidean distance in the PC subspace for two conformations is taken as a measure of their structural difference. In Fig. 1, 240,000 conformations form a cone shape in the three-dimensional structural space using the three most significant PCs (PC1-3). The folded conformations (with $C_\alpha\, RMSD \leq 2.2$Å compared with the NMR structure) are near the vertex of the cone in dark blue color and the unfolded conformations ($RMSD > 6.5$Å compared with the NMR structure) appear at the top of the cone in yellow and red colors.

To shorten the computational time required for the analysis of a network with a large number of nodes and edges, we can reduce the size of the network to a computationally tractable resolution by uniformly sampling from all the data. However, the transition regions which connect stable basins usually have low populations and deleting nodes uniformly will decrease the node density in the already sparsely sampled transition regions. We use a clustering technique to avoid this. Alternative methods have also been reported[79]. The clustering is done by combining neighboring nodes into a central node within a cutoff distance of $RMSD = 0.7$ Å. This corresponds to the

average conformational change in $2ps$ of the simulation at 465K. The central node of a cluster is chosen as the one with the largest number of neighbors within the cutoff distance. Combined neighboring nodes are removed and the central node is preserved. The population of the central node is increased by adding to it the population of the nodes which have been combined into the central node. Then we recalculate the number of neighbors for all the remaining nodes and pick the next central node for reduction. Clustering ends when there is no node that has neighbors within the cutoff distance. With this cutoff, we reduce the number of nodes to about $24,000$. The reduced network preserves the total population and the population distribution, but the resolution of the network is reduced. By choosing different cutoff distances, the network can be reduced to the desired resolution.

The properties of the kinetic network depend on how the nodes are connected and what the transition rates are between each pair of connected nodes. Pairs of nodes are connected if their distance is smaller than a cutoff distance ($RMSD = 1.8$Å). This cutoff is equal to the average conformational change after $20ps$ at 465K in a MD simulation of Trp-Cage (Simulation details are described in Appendix). The topology and kinetics of the resulting network does not strongly depend on this cutoff. Nodes which are disconnected from the largest subnetwork are removed to assure ergodicity of the network. More than 95% of nodes remain with this cutoff. To preserve the equilibrium population of each connected node, we assign microscopic rates to the edges that satisfy detailed balance conditions[49]

$$k_{ij} = \begin{cases} \mu_{ij}, & if\ p_{eq}(j) \leq p_{eq}(i) \\ \frac{p_{eq}(j)}{p_{eq}(i)}\mu_{ij}, & else \end{cases} \tag{1}$$

where k_{ij} is the Markovian rate from node j to i at a target temperature T_0, $p_{eq}(i)$ and $p_{eq}(j)$ are the equilibrium populations of the two nodes at T_0. $\mu_{ij} = \mu_{ji}$ is a base rate to be determined for each pair of nodes i and j and it sets the time scale of the network at T_0. μ_{ij} for the same pair of nodes may be chosen to vary with the temperature. In order to compare the kinetics of the network across multiple temperatures, μ_{ij} can in principle be calibrated by running multiple MD trajectories at different temperatures[49]. In this work, we focus on the flux distribution among pathways rather than determining the absolute value of the rate constants. Still we can make a rough estimation of the folding time scale in our model. As mentioned earlier in this section, the cutoff distance connecting two similar conformations in the network equals the average conformational change in $20ps$ calculated from the MD data at 465K. It is reasonable to set the base rate μ_{ij} between two connected conformations to $(20ps)^{-1}$. This assigns a real time unit to the network and the kinetics of the network can then be compared to experimental results.

2.2 Transition Path Theory (TPT) on the network: P_{fold}, flux and pathway analysis

To study the folding transitions on the network, the nodes are grouped into three subsets. The folded macrostate F and the unfolded macrostate U are defined as $RMSD \leq 2.2$Å and $RMSD \geq 6.5$Å, respectively. All the remaining nodes are assigned to the intermediate macrostate I. Transition Path Theory (TPT)[43,44,54] can then be applied to the network to extract pathways and their fluxes via the reactive flux $J_{i\to j}$ for transitions from node i to j, which is a function of the committor

probability P_{fold}[46] for node i and j

$$J_{i\to j} = \begin{cases} k_{ji}p_{eq}(i)[P_{fold}(j) - P_{fold}(i)], & if\ P_{fold}(j) > P_{fold}(i) \\ 0 & else \end{cases} \tag{2}$$

The P_{fold} for nodes belonging to macrostate I can be calculated by solving a set of linear equations based on the Markovian rate k_{ij}[80].

$$P_{fold}(i) = \sum_{i\neq j} P_{fold}(j)\frac{k_{ji}}{\sum_{l\neq i} k_{li}} \tag{3}$$

with boundary condition $P_{fold}(i) = 1$, $i \in F$ and $P_{fold}(i) = 0$, $i \in U$. It should be noted that P_{fold} values can also be well approximated by the eigenvectors of the transition rate matrix[81,82]. The obtained P_{fold} values are substituted into Eq.2 to yield the reactive flux $J_{i\to j}$ between all pairs of neighboring nodes. The total flux J leaving macrostate U equals the total flux entering macrostate F

$$J = \sum_{i\in U,\, j\notin U} J_{i\to j} = \sum_{j\notin F,\, i\in F} J_{j\to i} \tag{4}$$

The total flux J is the average number of observed $U \to F$ transitions per unit time. The folding rate constant k_f can be calculated from J

$$k_f = \frac{J}{\pi_U} \tag{5}$$

where $\pi_U = \sum_i^N p_{eq}(i)[1 - P_{fold}(i)]$ is the population of the system that contributes to folding transitions[83]. Note that although the sum is taken over all N nodes, only nodes in the I and U states contribute to π_U since their $P_{fold} < 1$. In the two-state limit where $p_{eq}(I) \ll p_{eq}(U) + p_{eq}(F)$, $\pi_U \approx p_{eq}(U)$ and Eq.5 reduces to $k_f = J/p_{eq}(U)$.

The total flux can be decomposed into individual pathways. A folding pathway in this context is defined as a series of hops between pairs of connected nodes on the network, starting from the macrostate U and ending at the macrostate F, moving monotonically from low P_{fold} value nodes to high P_{fold} value nodes. A pathway is different from a trajectory on the network in that a pathway does not include all of the recrossings between nodes that do not advance P_{fold} values. In the sense of providing information about the mechanism of a transition, a pathway is a more concise description than a trajectory, but a trajectory does provide more detailed kinetic information. There is no unique way to extract pathways from the network. One way of particular interest, as described in ref. [43], is to repeatedly find and remove the pathway with the largest flux from the network until there is no reaction pathway left. The flux of each extracted pathway is the minimum of all the reactive fluxes between adjacent nodes along the pathway. For a different target temperature, we create a new network with a different equilibrium population distribution without further simulations. The topology of the network stays the same, but the equilibrium population of each node $p_{eq}(i)$, the Markovian rate k_{ij} for each edge and $P_{fold}(i)$ for each node in the macrostate I are recalculated for each temperature. A new set of folding pathways with new fluxes can then be extracted at the new target temperature.

In order to reveal mechanistic insights into the folding transition, we need to define a way to group the large number of pathways generated from the TPT decomposition algorithm into a much

smaller number of clusters of pathways which we call tubes. One way to quantify the difference between two pathways is to calculate the distances between them at different stages of folding and take the average of all the distances. We use P_{fold} values as the measure for different stages of folding. Each pathway is divided into N segments evenly via their P_{fold} values (from 0 to 1) along the pathway. One node from each P_{fold} interval on a pathway is chosen for the calculation of the pair distances between the two pathways. We define the Root-Mean-Square distance (RMS) between two pathways among the N pairs of nodes with the same P_{fold} values as:

$$RMS(path^A, path^B) = \sqrt{\frac{1}{N}\sum_{i}^{N}[distance(n^A_{P^i_{fold}}, n^B_{P^i_{fold}})]^2} \tag{6}$$

where $n^A_{P^i_{fold}}$ means the node with $P_{fold} = P^i_{fold}$ on pathway A. $n^B_{P^i_{fold}}$ is defined in the same way for the nodes on pathway B. Twelve different P_{fold} values, evenly distributed between 0 and 1, are chosen for $\{P^i_{fold}\}$. With this definition of the distance between any two pathways, standard clustering techniques can be applied.

2.3 Stochastic simulations on the network: P_{fold}, flux and folding trajectories

Stochastic simulations on the network is another powerful tool to explore the thermodynamic and kinetic properties of the network. Compared with the computational cost of obtaining reactive trajectories from all-atom simulations, there is a much lower computational burden to collect reactive trajectories on the network because all the conformations are precalculated and the potential energies and forces do not need to be evaluated again. Stochastic simulations on the network not only serve as a benchmark for the TPT calculation to test the validity of the flux formulas but also provide more kinetic information than that is contained in the pathways. However, the TPT analysis indeed provides a framework for analyzing the stochastic trajectories as we discuss below. We run stochastic simulations using the Gillespie algorithm[84] on the network to construct folding trajectories and also obtain thermodynamic and kinetic properties. For example, to estimate the P_{fold} value of a node i, we start numerous simulations from node i and stop each simulation when either the macrostate F or U is reached. The fraction of trajectories which reach the macrostate F before reaching U gives the estimate of P_{fold}. We can also start a simulation from a random node and run a long trajectory that contains a large number of folding/unfolding events. The equilibrium population of each node is proportional to the total residence time at that node during the simulation. The folding flux can be estimated by

$$J = \frac{\#\,of\ folding\ events}{Total\ simulation\ time} \tag{7}$$

We can determine whether there is a good correspondence between the folding trajectories generated from simulations and the pathways extracted using TPT. In order to do this, we need to assign trajectories to pathways. To determine whether a trajectory belongs to any of the pathways, we use the RMS defined in Sec.2.2. If $RMS(trj^i, path^j) = min\{RMS(trj^i, path^k)\}, k = 1, ..., N_{path}$ and $RMS(trj^i, path^j) \leq cutoff$, we define that trajectory i belongs to pathway j. Here the $cutoff$ is set as the minimum RMS between any two pathways. If $min\{RMS(trj^i, path^k)\} > cutoff$,

we define the trajectory as not belonging to any extracted pathways. Note that to calculate the RMS between a trajectory and a pathway, the trajectory has to be rearranged in the order of increasing P_{fold} values. More specifically, if some part of the trajectory visits the same node twice (recrossing), all the other nodes visited in between the recrossing are excluded from the RMS calculation. We can then compare the stochastic trajectories with the TPT pathways. Moreover, some kinetic properties that are difficult to calculate using TPT can be obtained from the stochastic simulations. For example, a pathway extracted using the TPT algorithm is associated with a large number of stochastic trajectories. The information about the first passage times (FPT) for all of these trajectories is not contained in the flux of that folding pathway. But we can record FPT statistics from a long trajectory which contains numerous transition events. The distribution of the FPT can then be used to test Trp-Cage's two-state behavior as well as calculate the folding rate constant.

3 Results and discussion

The first 20ns of the REMD simulation is excluded from the analysis described in this section. 240,000 conformations are collected from the rest of the simulation data (20-50ns) with 15,000 samples at each of the 16 temperatures. The melting curve is obtained by using the distance from the NMR structure (PDB ID 1L2Y) as a measure to distinguish the folded and unfolded macrostates. $C_\alpha\, RMSD \leq 2.2$Å is defined as the folded state. As shown in the left plot of Fig.2, the folding temperature is around 450K. The high folding temperature is likely an artifact from the OPLS force field[85] or the implicit solvent model[86]. A similar folding temperature was also observed in all-atom simulations that use Amberff94 force field and explicit solvent model[67,69]. As suggested in previous work by Wang et al.[87], the kinetic pathways undergo a transition from few paths to many paths above the folding temperature for fast-fold proteins. We choose 465K, which has a folded population of 0.42, as a high temperature reference for further analysis of the kinetics. On the other hand, we choose 363K as a low temperature reference for kinetic analysis because at this temperature we can obtain stochastic simulation results with much better statistics. At lower temperatures the trajectories tend to be trapped in local minima. The fraction of folded conformations is very similar at 363K and 298K. The distributions of distance from the NMR structure at the two temperatures are plotted in the right plot of Fig.2. The $RMSD$ distributions show a high peak around the folded state at 363K and two well-separated peaks at 465K. Both plots are similar to a previous report[69], except that our $RMSD$ distribution data suggests two-state behavior for the system near the folding temperature.

Another important piece of information which couples thermodynamics with structure is the potential of mean force (PMF) projected along different principal components (PC). We investigate the correlations between the three most significant PCs and the C_α internal coordinates to reveal more structural insights. PC1 has strong correlations with the internal coordinates which describe the $\alpha-$helix (residues 2-8), the $3_{10}-$helix (residues 11-14) and the distance between the C-terminus and N-terminus. PC2 involves those internal coordinates in the turn region (residues 8-11), the $3_{10}-$helix and the distance between the termini, similar to PC1 but it does not involve internal coordinates of the $\alpha-$helix. PC3 is similar to PC2 but excludes the contribution from the end to end distance of the polypeptide chain. Fig. 1 suggests that PC1 separates well unfolded from folded structures and indicates how folding progresses, PC2 and PC3 show the decreasing variations on

the two components as the structure approaches the folded state. Previous work on Trp-Cage by Juraszek and Bolhuis[88] has shown that the combination of $RMSD$ and $RMSD_{helix}$ ($RMSD$ from residue 2 to residue 8) is one of the best choices for the reaction coordinate of Trp-Cage. The correlation coefficient between PC1 and $RMSD$ is 0.94 and the correlation coefficient between PC1 and $RMSD_{helix}$ is 0.87.

The PMF of Trp-Cage along PC1 and PC2 is plotted at two different temperatures: 363K and 465K (Fig3). At each temperature, the PMF is calculated in two ways. One is to use the simulation data only from the target temperature, each sample has the same weight as the others (the right two plots). The other takes all the simulation data, calculates the corresponding WHAM weight[49,77] for each sample and obtains the PMF (the left two plots). WHAM gives a more complete PMF by taking advantage of all the simulation data. This is especially evident for the low temperature. Most of the 363K conformations are near the folded state because of the high stability of the mini-protein at low temperature[55]. By reweighting the conformations from higher temperatures, we are able to construct a much more complete PMF at 363K. Above the folding temperature, however, the two PMFs look more similar. The folded state ($RMSD \leq 2.2$Å) is located inside the red circle in the plot. The unfolded state is inside the blue ellipse. Note that the landscape looks much smoother at the higher temperature, i.e. the variation in the scale from low to high energy region is much smaller.

To obtain kinetic properties of the network, both stochastic simulations and transition path theory (TPT) can be applied. We focus on the temperature 465K to test the correspondence between pathways constructed using the TPT algorithm and stochastic trajectories obtained by Gillespie simulations. First, we note that the population of each node determined from the stochastic simulations successfully reproduces the WHAM weight for the node. As shown in the left plot of Fig.4, the deviation only becomes noticeable for extremely low populated nodes. To estimate P_{fold} of the nodes in the I state, we simulate 100 trajectories from each node and calculate the fraction of folding trajectories. The comparison of P_{fold} calculated by solving the linear equations (Eq.3) and that from the Gillespie simulations is shown in the right plot of Fig.4. They agree very well. The total folding flux and the folding rate constant obtained from the simulations also agree very well with the results of TPT, as shown in Table 1. We extract 10,000 folding pathways at 465K using the TPT algorithm and then group them into 33 different folding tubes by clustering the pathways. The flux of a folding tube is the sum of the fluxes of all pathways that belong to the tube. The probability that a stochastic folding trajectory belongs to a certain folding tube should be proportional to the flux of the tube. To test this, we randomly generated 1000 folding trajectories from simulations and assigned them to one of the folding tubes as described in Methods Sec.2.3. As shown in Fig.5, the fraction of stochastic trajectories that belong to a particular folding tube agrees well with the flux of that tube expressed as a fraction of the total flux from all the tubes. The first passage time (FPT) distribution of the 1000 trajectories fits well with a single exponential curve ($R^2 > 0.998$). This coincides with the two-state kinetics observed previously in the experiment of Trp-Cage folding[57]. However, further investigation is needed to elucidate the apparent paradox that so many different folding routes together exhibit effective two-state kinetics.

We have investigated the temperature dependence of the ensemble of folding pathways. Thousands of different pathways are extracted using the TPT algorithm at 363K and 465K as shown in the left plot of Fig.6. To analyze these pathways, they are clustered using the pathway RMS criterion introduced in section2.2. There are many more routes at 465K (33 tubes) than at 363K (3 tubes). The right plot of Fig.6 shows the flux distribution among the tubes. The three folding

tubes A, B and C from 363K are also observed at 465K, but each with an altered flux. In Fig.7, the centroid pathway of each tube is projected onto the two-dimensional PMF. Folding tubes A, B and C are also labeled in the plot. The thickness of the line corresponds to the flux of each tube. Folding pathways are very localized at low temperature due to the roughness of the landscape, while they show much more variety above the folding temperature. In addition to the PMFs in Fig.3, this is further evidence that the free energy landscape is much smoother at high temperatures, so that many new pathways that are forbidden at low temperatures become available.

One interesting observation about the pathways at 465K is that the folding route which the system takes to fold highly depends on the structure of the unfolded state it starts from. There is a strong correlation (0.90) between the heterogeneity inside the unfolded state and that of the pathways, i.e., the distance between two pathways is strongly correlated with the structural difference between the unfolded basins from which the folding is initiated. This observation is suggestive of the kinetic hub folding model recently proposed by Pande et al.[89] for which there are much faster transitions from different unfolded basins to the native state than among the unfolded basins. Further analysis of possible connections between the kinetic hub folding model and the Trp-Cage folding paths observed in our kinetic network model is in progress.

We report the conformational details of two major folding tubes in Fig.8. In folding tube A, seen at both 363K and 465K but with much higher flux at the low temperature (Fig.6), the chain starts with a compact structure with the $Trp6$ residue packed between $Tyr3$ and $Pro17$ and no helical content. As the helix forms, $Trp6$ is trapped and forms several non-native contacts, especially with $Arg16$. In order to obtain the correct packing for the core, the helix has to dissolve and this provides more room for $Trp6$ to break the contact with $Arg16$ and leads to the correct packing of the core. For this pathway, the analysis shows that to reach the correct packing of the hydrophobic core, the helix has to form, break and reform. In folding tube B, also observed at both temperatures but with increased flux at higher temperature, the helix forms first while the chain is still extended. Then close contacts are formed among $Tyr3$, $Trp6$ and $Pro12$. Finally the residues near the N-terminus ($Pro17-19$) wrap around $Trp6$ to complete the cage. The existence of the two dominant folding paths is consistent with the fact that both quick collapse of the core[65] and helical structure in the denatured state[64] have been observed in experiments. These two folding routes are also qualitatively similar to the report by Juraszek et al.[66]. However our more detailed pathway analysis suggests that the formation of helix and hydrophobic core is a more complex process than captured by a sequential model.

4 Conclusions

We present a new approach to study protein folding pathways and folding mechanisms for mini-proteins and apply our kinetic network model to study the folding of Trp-Cage. We use REMD with a large range of temperatures to sample the conformational space. The conformations collected from all temperatures form a discretized state space. The equilibrium population for each discrete state at a target temperature can be calculated using WHAM. By connecting states with similar structures and creating edges that satisfy detailed balance conditions, we construct a kinetic network that preserves the equilibrium population distribution of the state space. The network contains new pathways that were not explicitly realized in the original REMD trajectories in the continuous conformational space. Folding pathways can be extracted from the network using transition path

theory (TPT) or generated from stochastic simulations on the network at multiple temperatures. Above the folding temperature, a large number of diverse folding pathways "flood" the energy landscape with comparable fluxes. At low temperature, however, only one or two dominating folding routes that lead to the native state are observed. The analysis of folding fluxes based on stochastic simulations on the network confirms the transition path theory predictions, as shown in Table 1 and Fig.5.

Our approach to the protein folding problem uses replica exchange all-atom simulations to construct a discretized network together with TPT tools and stochastic simulations to interrogate the network. It can be applied to other problems in protein biophysics for which long time scale dynamics is a central issue. In a forthcoming report we apply these tools to analyze transition paths associated with ligand binding to a protein in its native state[90].

5 Acknowledgments

We thank Wei Dai for his assistance with analysis of enhanced sampling with REMD compared with MD presented in the appendix. This work is supported by a grant from the National Institutes of Health (GM305080). The calculations reported in this work have been performed at the BioMaPS High Performance Computing Center at Rutgers University funded in part by the NIH shared instrumentation grant no. 1 S10 RR022375.

References

[1] Onuchic, J.; Wolynes, P. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70–75.

[2] Kolinski, A.; Skolnick, J. *Polymer* **2004**, *45*, 511–524.

[3] Onuchic, J.; Luth-Schulten, Z.; Wolynes, P. *Annu.Rev.Phys.Chem.* **1997**, *48*, 545–600.

[4] Dill, K. *Protein Sci.* **1999**, *8*, 1166–1180.

[5] Kolinski, A.; Ilkowski, B.; Skolnick, J. *Biophys. J.* **1999**, *77*, 2942–2952.

[6] Kilmov, D.; Thirumalai, D. *Proc.Natl.Acad.Sci.USA* **2000**, *97*, 2544–2549.

[7] Dokholyan, N.; Buldyrev, S.; Stanley, H. E. *J.Mol.Bio.* **2000**, *296*, 1183–1188.

[8] Ozkan, S.; Dill, K.; Bahar, I. *Protein Sci.* **2002**, *11*, 1958–1970.

[9] Ozkan, S.; Dill, K.; Bahar, I. *Biopolymers* **2003**, *68*, 35–46.

[10] Klimov, D. K.; Thirumalai, D. *J.Mol.Biol.* **2005**, *353*, 1171–1186.

[11] Dinner, A.; Lazaridis, T.; Karplus, M. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 9068–9073.

[12] Pande, V.; Rokhsar, D. *Proc.Natl.Acad.Sci. USA* **1999**, *96*, 9062–9067.

[13] Ma, B.; Nussinov, R. *J.Mol.Bio.* **2000**, *296*, 1091–1104.

[14] Garcia, A.; Sanbonmatsu, K. *Proteins* **2001**, *42*, 345–354.

[15] Zagrovic, B.; Sorin, E.; Pande, V. *J. Mol. Bio.* **2001**, *313*, 151–169.

[16] Zhou, R.; Berne, B.; Germain, R. *Proc.Natl.Acad.Sci. USA* **2001**, *98*, 14931–14936.

[17] Brooks III, C. L. *Acc.Chem.Res.* **2002**, *35*, 447–454.

[18] Pande, V.; Baker, I.; J., C.; Elmer, S.; Khaliq, S., S. Larson; Rhee, Y.; Shirts, M.; Snow, C.; Sorin, E.; Zagrovic, B. *Biopolymers* **2003**, *68*, 91–109.

[19] Bolhuis, P. *Proc.Natl.Acad.Sci. USA* **2003**, *100*, 12129–12134.

[20] Felts, A.; Harano, Y.; Gallicchio, E.; Levy, R. *Proteins* **2004**, *56*, 310–321.

[21] Shaw, D. E. et al. *Communications of the ACM* **2008**, *51*, 91.

[22] Beck, D. *Methods* **2004**, *34*, 112–120.

[23] Ferguson, N.; Day, R.; Johnson, C. M.; Allen, M. D.; Daggett, V.; Fersht, A. R. *Journal of Molecular Biology* **2005**, *347*, 855–870.

[24] Schumaker, M. F.; Pomès, R.; Roux, B. *Biophys. J.* **2000**, *79*, 2840–2857.

[25] Hummer, G.; Kevrekidis, I. G. *J. Chem. Phys.* **2003**, *118*, 10762.

[26] Kopelevich, D. I.; Panagiotopoulos, A. Z.; Kevrekidis, I. G. *J. Chem. Phys.* **2005**, *122*, 044908.

[27] Best, R. B.; Hummer, G. *Phys. Rev. Lett.* **2006**, *96*, 228104.

[28] Yang, S.; Onuchic, J. N.; Levine, H. *J. Chem. Phys.* **2006**, *125*, 054910.

[29] Yang, S.; Onuchic, J. N.; Garcìa, A. E.; Levine, H. *J. Mol. Biol.* **2007**, *372*, 756–763.

[30] Chekmarev, D. S.; Ishida, T.; Levy, R. M. *J. Phys. Chem. B* **2004**, *108*, 19487–19495.

[31] Rogal, J.; Bolhuis, P. *J.Chem.Phys.* **2008**, *129*, 224107.

[32] Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299–306.

[33] Apaydin, M. S.; Brutlag, D. L.; Guestrin, C.; Hsu, D.; Latombe, J.-C. *Proceedings of RECOMB '02* **2002**, 12–21.

[34] Hubner, I. A.; Deeds, E. J.; Shakhnovich, E. I. *Proceedings of the National Academy of Sciences* **2006**, *103*, 17747–17752.

[35] Ozkan, S. B.; Dill, K. A.; Bahar, I. *Protein Sci.* **2002**, *11*, 1958–1970.

[36] Ye, Y.-J.; Ripoll, D. R.; Scheraga, H. A. *Comp. Theor. Polymer Sci.* **1999**, *9*, 359–370.

[37] Singhal, N.; Snow, C. D.; Pande, V. S. *J. Chem. Phys.* **2004**, *121*, 415–425.

[38] Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571–6581.

[39] 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.

[40] Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299–306.

[41] Huang, X.; Bowman, G. R.; Bacallado, S.; Pande, V. S. *Proceedings of the National Academy of Sciences* **2009**, *106*, 19765–19769.

[42] Buchete, N.-V.; Hummer, G. *Phys. Rev. E* **2008**, *77*, 030902.

[43] Metzner, P.; Schütte, C.; Vanden-Eijnden, E. *Multiscale Model. Simul.* **2009**, *7*, 1192–1219.

[44] Berezhkovskii, A.; Hummer, G.; Szabo, A. *J.Chem.Phys.* **2009**, *130*, 205102.

[45] Noe, F.; Schutte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proceedings of the National Academy of Sciences* **2009**, *106*, 19011–19016.

[46] Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; S.Shakhnovich, E. *J.Chem.Phys.* **1998**, *108*, 334–350.

[47] Gillespie, D. T. *J. Phys. Chem.* **1977**, *81*, 2340–2361.

[48] Andrec, M.; Felts, A. K.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6801–6806.

[49] Zheng, W.; Andrec, M.; Gallicchio, E.; Levy, R. M. *J.Phys.Chem.B* **2009**, *113*, 11702–11709.

[50] Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

[51] Zheng, W.; Andrec, M.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 15340–15345.

[52] van der Spoel, D.; Seibert, M. *Phys.Rev.Lett.* **2006**, *96*, 238102.

[53] Andrec, M.; Felts, A. K.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6801–6806.

[54] Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 19011–19016.

[55] Neidigh, J. W.; Fesinmeyer, M. R.; Andersen, N. H. *Nature Structural Biology* **2002**, *9*, 425–430.

[56] Simmerling, C.; Strockbine, B.; Roitberg, A. E. *Journal of the American Chemical Society* **2002**, *124*, 11258–11259.

[57] Qiu, L.; Pabit, S. A.; Roitberg, A. E.; Hagen, S. J. *Journal of the American Chemical Society* **2002**, *124*, 12952–12953.

[58] Snow, C. D.; Zagrovic, B.; Pande, V. S. *Journal of the American Chemical Society* **2002**, *124*, 14548–14549, PMID: 12465960.

[59] Chowdhury, S.; Lee, M. C.; Xiong, G.; Duan, Y. *Journal of Molecular Biology* **2003**, *327*, 711–717.

[60] Pitera, J. W. *Proceedings of the National Academy of Sciences* **2003**, *100*, 7587–7592.

[61] Zhou, R. *Proceedings of the National Academy of Sciences* **2003**, *100*, 13280–13285.

[62] Linhananta, A.; Boer, J.; MacKay, I. *The Journal of Chemical Physics* **2005**, *122*, 114901.

[63] Ding, F.; Buldyrev, S.; Dokholyan, N. *Biophysical Journal* **2005**, *88*, 147–155.

[64] Ahmed, Z.; Beta, I. A.; Mikhonin, A. V.; Asher, S. A. *Journal of the American Chemical Society* **2005**, *127*, 10943–10950.

[65] Neuweiler, H. *Proceedings of the National Academy of Sciences* **2005**, *102*, 16650–16655.

[66] Juraszek, J.; Bolhuis, P. G. *Proceedings of the National Academy of Sciences* **2006**, *103*, 15859–15864.

[67] Paschek, D.; Nymeyer, H.; Garc\'ia, A. E. *Journal of Structural Biology* **2007**, *157*, 524–533.

[68] Juraszek, J.; Bolhuis, P. *Biophysical Journal* **2008**, *95*, 4246–4257.

[69] Paschek, D.; Hempel, S.; Garcia, A. E. *Proceedings of the National Academy of Sciences* **2008**, *105*, 17754–17759.

[70] Marinelli, F.; Pietrucci, F.; Laio, A.; Piana, S. *PLoS Computational Biology* **2009**, *5*, e1000452.

[71] Day, R.; Paschek, D.; Garcia, A. E. *Proteins: Structure, Function, and Bioinformatics* **2010**, 1889–1899.

[72] Ota, M. *Proceedings of the National Academy of Sciences* **2004**, *101*, 17658–17663.

[73] Chowdhury, S.; Lee, M. C.; Duan, Y. *The Journal of Physical Chemistry B* **2004**, *108*, 13855–13865.

[74] Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

[75] Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479–499.

[76] Kumar, S.; Bouzida, D.; Swendsen, R.; A., K. P.; Rosenberg, J. M. *J.Comp.Chem.* **1992**, *13*, 1011–1021.

[77] Gallicchio, E. .; Andrec, M. .; Felts, A.; Levy, R. *J.Phys.Chem.B* **2005**, *109*, 6722–6731.

[78] Jolliffe, I. T. *Principal Component Analysis* **2002**, *Springer*, New York.

[79] Huang, X.; Yao, Y.; Bowman, G. R.; Sun, J.; Guibas, L. J.; Carlsson, G.; Pande, V. S. *Pasific Symposium on Biocomputing* **2010**, *15*, 228–239.

[80] Singhal, N.; Snow, C. D.; Pande, V. *J.Chem.Phys.* **2004**, *121*, 415–425.

[81] Berezhkovskii, A.; Szabo, A. *The Journal of Chemical Physics* **2004**, *121*, 9186.

[82] Buchete, N.; Hummer, G. *The Journal of Physical Chemistry B* **2008**, *112*, 6057–6069.

[83] Vanden-Eijnden, E.; Venturoli, M. *The Journal of Chemical Physics* **2009**, *131*, 044120.

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

[85] Zhou, R. *Proceedings of the National Academy of Sciences* **2001**, *98*, 14931–14936.

[86] Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. *Proteins* **2004**, *56*, 310–321.

[87] Wang, J.; Onuchic, J.; Wolynes, P. *Physical Review Letters* **1996**, *76*, 4861–4864.

[88] Juraszek, J.; Bolhuis, P. G. *Biophys.J.* **2008**, *95*, 4246–4257.

[89] Bowman, G. R.; Pande, V. S. *Proceedings of the National Academy of Sciences* **2010**, *107*, 10890–10895.

[90] Deng, N.; Zheng, W.; Gallicchio, E.; Levy, R. M. *to be submitted* **2010**,

[91] Rosta, E.; Hummer, G. *The Journal of Chemical Physics* **2009**, *131*, 165102.

Table 1: The correspondence between the kinetic properties obtained by stochastic simulations and those constructed using Transition Path Theory (TPT) at 465K.

	Stochastic Simulation	Transition Path Theory
Total flux J (μs^{-1})	2.3	2.3
Folding rate k_f (μs^{-1})	*10.1	9.5

*We compared the folding rate constant k_f obtained from our model with the experimental value[57] at the folding temperature (465K for the model and 317K for the experiment). Our calculation is one order of magnitude larger than the experimental value ($10.1\mu s^{-1}$ vs. $0.6\mu s^{-1}$). Considering the fact that the folding temperature is significantly higher in the simulation and there is less friction using the implicit solvent model, our model does give a reasonable estimate of the real folding time scale.

For the stochastic simulation approach, the total flux is calculated using Eq.7 and the folding rate constant equals the inverse of the mean first passage time (MFPT) of the stochastic trajectories. For the TPT approach, the total flux is calculated using Eq.4 and the folding rate constant is calculated using Eq.5.

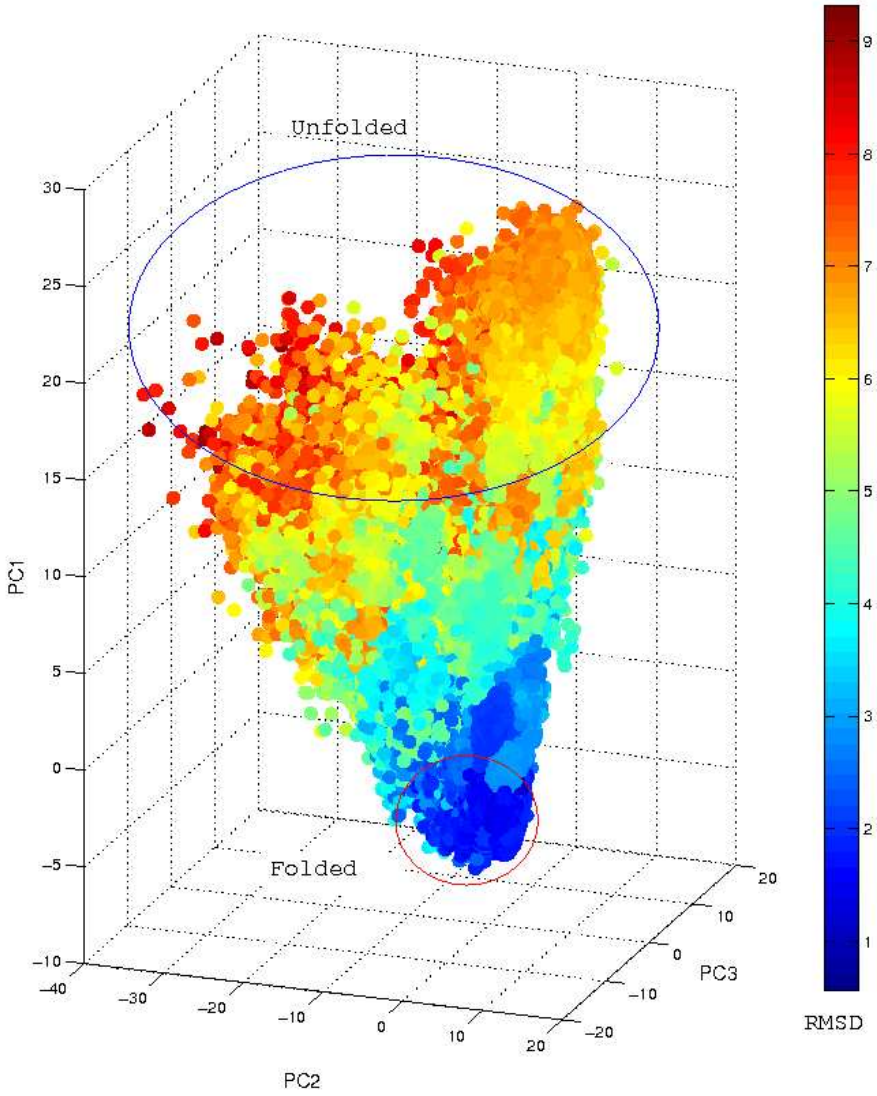


Figure 1: 240,000 conformations are plotted as points in the principal component (PC) subspace using the first three most significant PCs. Each PC is a linear combination of 54 $C_\alpha - C_\alpha$internal coordinates. The points form a special cone shape. The color code corresponds to the $C_\alpha\, RMSD$ from the NMR structure. The folded state ($RMSD \leq 2.2$) is located near the vertex of the cone, while the unfolded state ($RMSD \geq 6.5$) are near the top of the cone. PC1 separates well the folded and unfolded states. PC2 and PC3 show the decreasing variations on the coordinates as folding progresses.

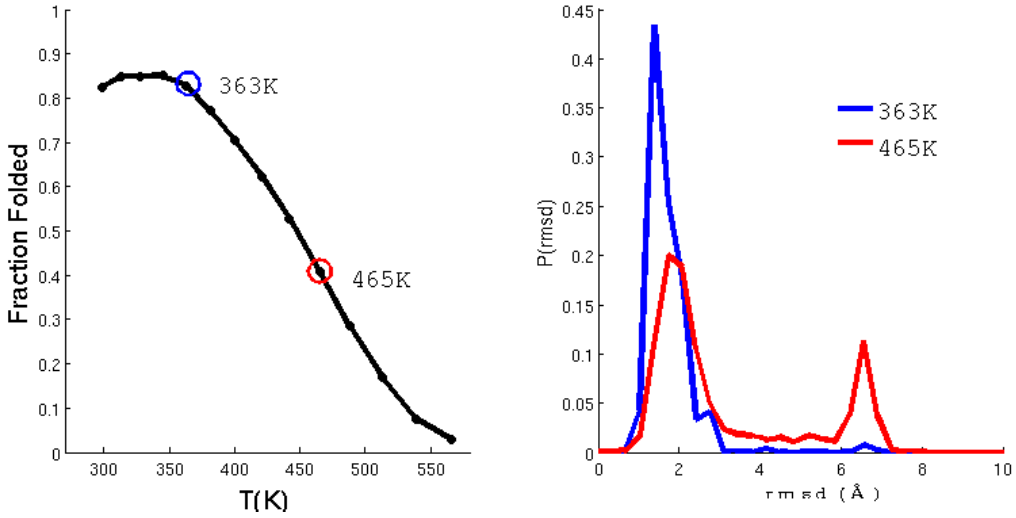


Figure 2: Conformational information collected from REMD simulations of Trp-Cage. (Left)Fraction of folded conformations ($C_\alpha\, RMSD \leq 2.2$) as a function of temperatures. The folding temperature is around 450K. (Right)The $RMSD$ distributions at 2 different temperatures: 363K and 465K.

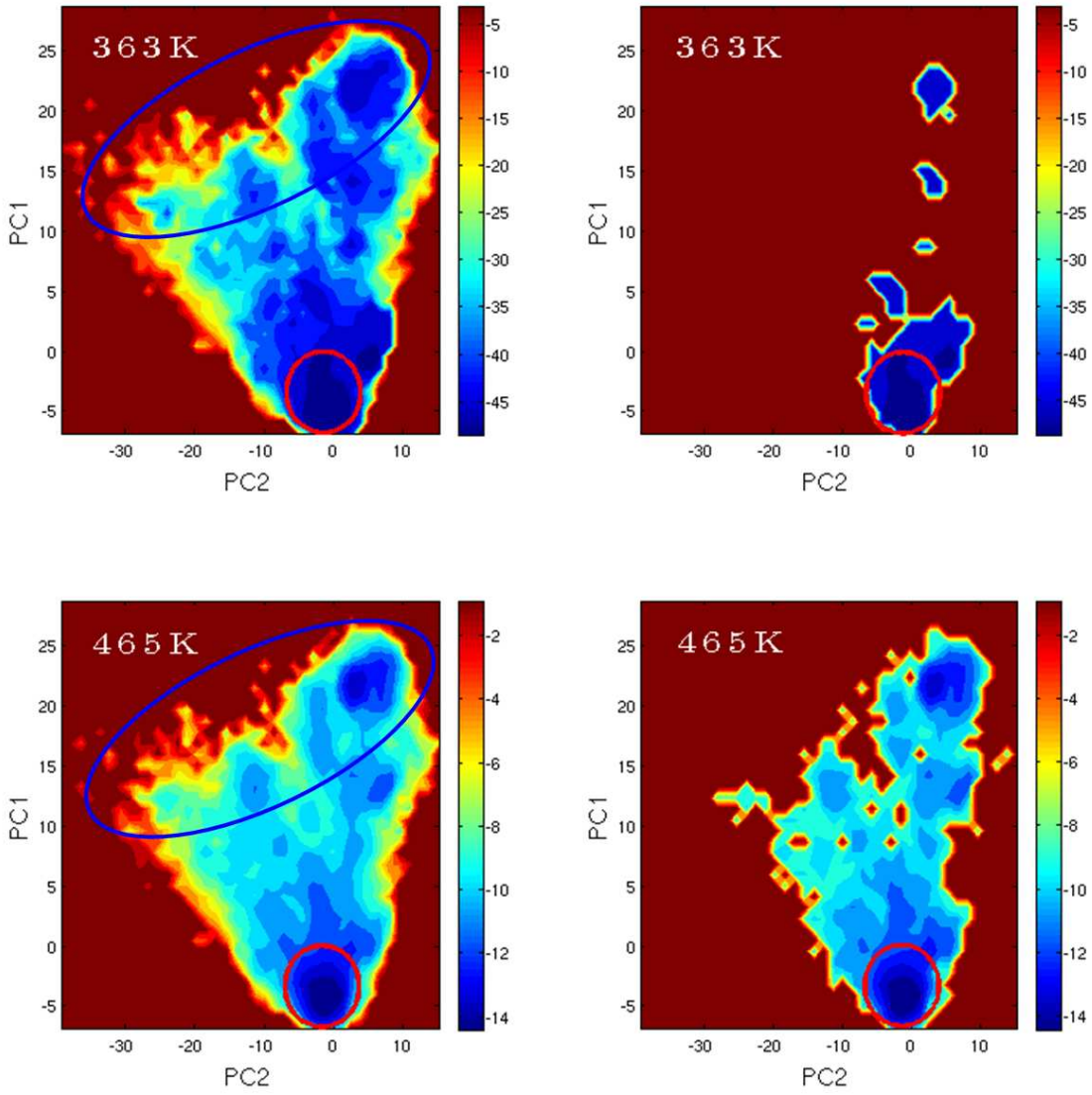


Figure 3: Potential of the mean force (PMF) of Trp-Cage at two temperatures: 363K and 465K. The left two plots use data from all 16 temperatures with calculated WHAM weights. The right two plots use only the data from that single temperature. WHAM gives a much more complete PMF, especially at 363K. The landscape is much smoother at 465K. The defined folded state ($C_\alpha\, RMSD \leq 2.2$Å) is located inside the the red circle. The unfolded state is within the blue circle.

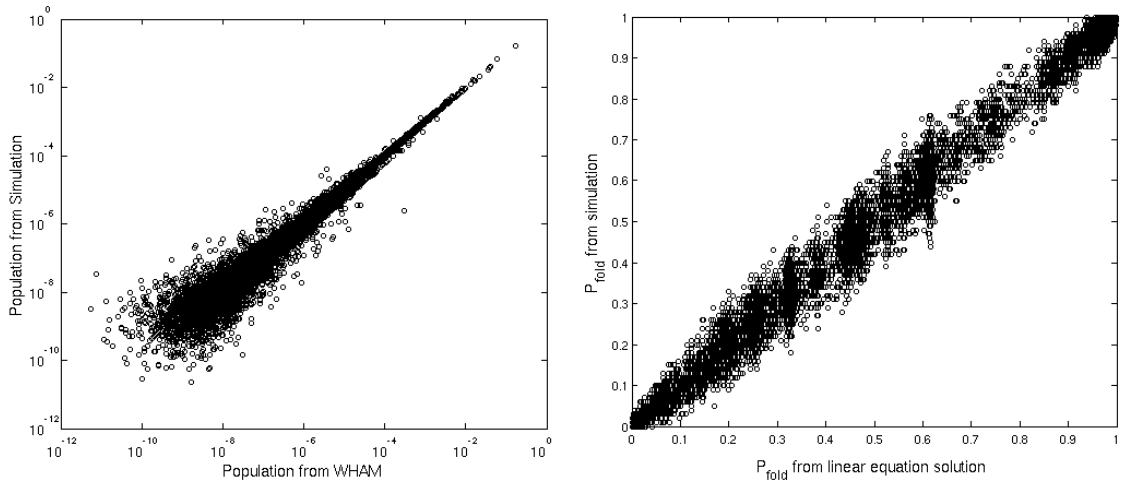


Figure 4: (Left)Comparisons of the equilibrium population of each node from stochastic simulations and its WHAM weight at 465K. Deviations only appear for very low populated nodes; (Right)Comparisons of the P_{fold} for each node obtained by solving the linear equations[80] and the P_{fold} from stochastic simulations at 465K. 100 folding/unfolding trajectories are simulated from each node and the fraction of folding trajectories gives an estimate for the P_{fold} of each node.

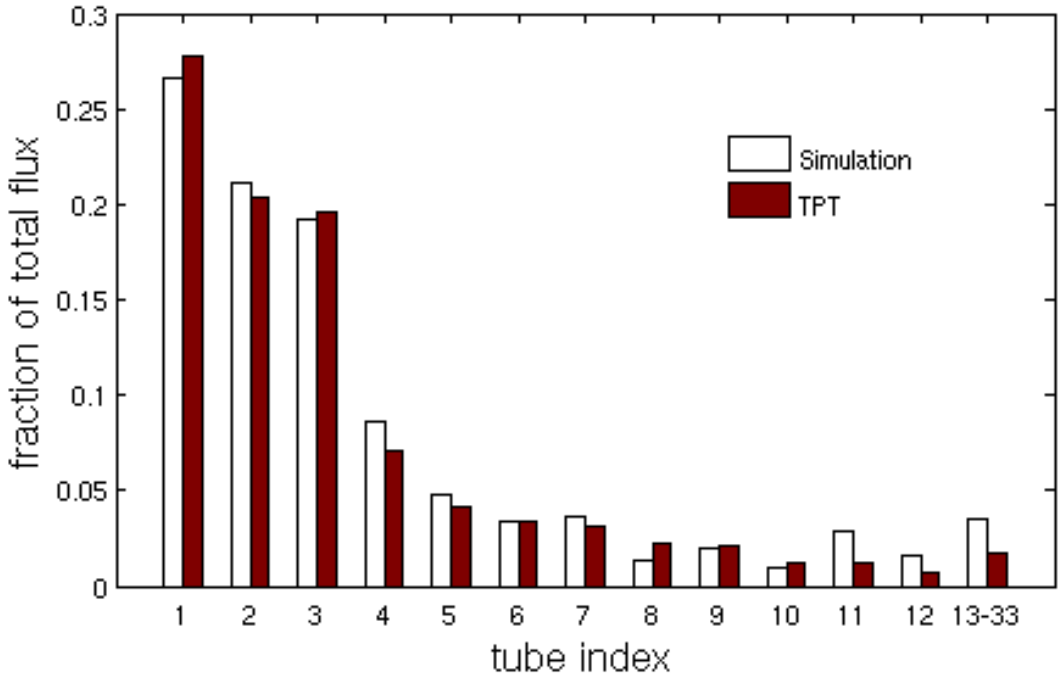


Figure 5: 1000 folding trajectories are generated from stochastic simulations on the network at 465K. Each trajectory can be assigned to a particular folding tube using methods described in 2.3. We compared the fraction of the total flux for each folding tube calculated from the TPT analysis and the fraction of simulation trajectories which belong to the same tube and found excellent agreement. Folding tubes with indices from 13 to 33 have relatively small fluxes. They are put into a single bin.

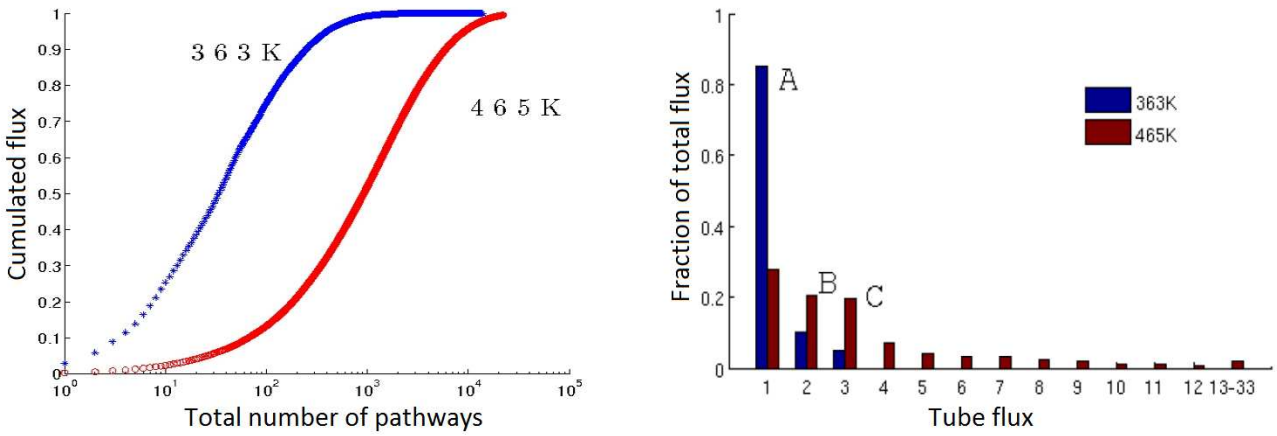


Figure 6: Thousands of different pathways are extracted using the TPT algorithm at two temperatures as shown in the left plot. By clustering the large number of pathways into folding tubes, a more physically meaningful picture of the diversity of folding routes emerges. The right plot is the flux distribution among the folding tubes. There are many more folding tubes at 465K (33 tubes) than at 363K (3 tubes). The 3 folding tubes at 363K, labeled with A, B and C, are also observed at the higher temperature, but with altered fluxes. At 465K, the total folding flux is spread into many more folding tubes that are not available at 363K.

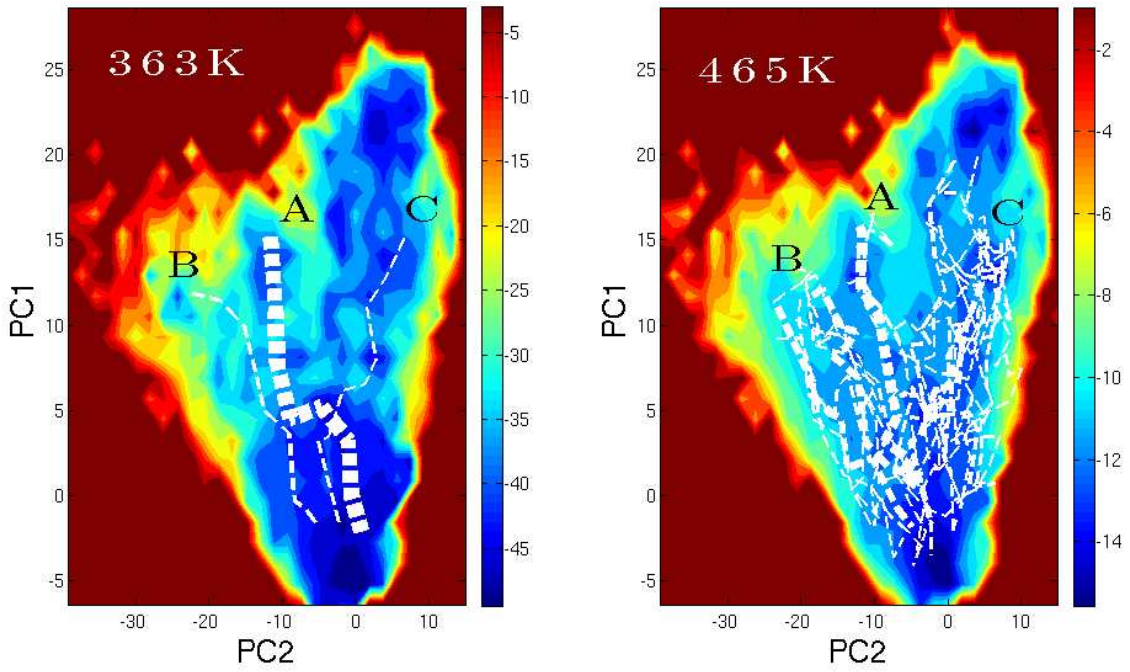


Figure 7: Tubes of folding pathways for Trp-Cage projected onto a two-dim PMF at two temperatures: 363K and 465K. There are 3 folding tubes at 363K and 33 folding tubes at 465K (all in dashed lines). The 3 folding tubes at 363K also occur at 465K (labeled with A, B and C), but with altered fluxes. The thickness of each lines stands for the total flux of that tube. Folding pathways are very localized at 363K due to the roughness of the landscape and are of much more variety and complexity at 465K. Each transition tube starts from the unfolded state and ends at the folded state.

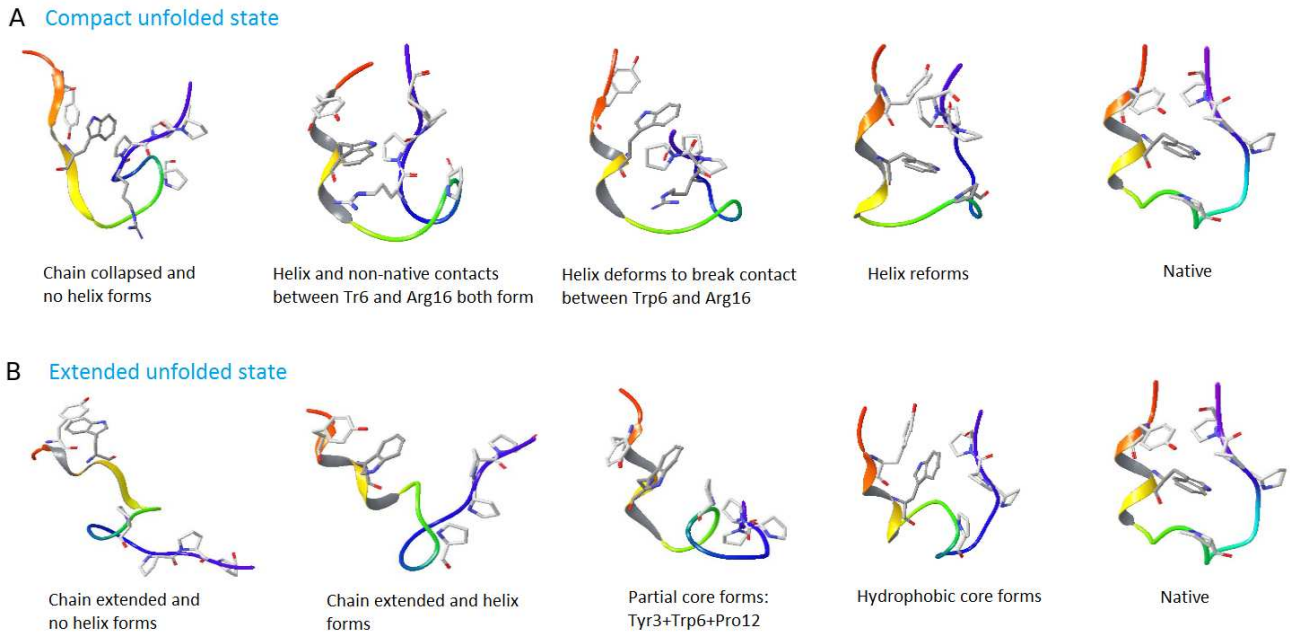


Figure 8: Two folding mechanisms from folding tube A and B are observed at both 363K and 465K. They start from two totally different unfolded structures: a compact conformation and an extended conformation. Side chains $Tyr3$, $Trp6$, $Pro12$, $Arg16$, $Pro17-19$ are shown using tube representation. (A)The folding route starts with a collapsed chain with no helical content. As helix forms, the core also forms one of its non-native contacts with $Arg16$. In order to obtain the correct packing for the core, helix dissolves and the contact between $Trp6$ and $Arg16$ breaks up. This leads to the correct packing of the core. (B)This folding route starts with an extended structure. Helix forms first while the chain is still extended. Then close contacts are formed among $Tyr3$, $Trp6$ and $Pro12$. Finally the residues near the N-terminus wrap around $Trp6$ to complete the cage. The two folding routes have different order in secondary and tertiary structure formation: one forms the core first, the other forms the helix first. However, our detailed pathway analysis suggests that the formation of helix and hydrophobic core is a more complex process than captured by a sequential model. For example, in folding route A both the breaking and reforming of helix occur on the way for Trp-cage to reach the final native structure.

6 Appendix: Enhanced sampling with REMD compared with MD

In this appendix, we discuss the benefit of using REMD for sampling the conformational space of Trp-Cage compared with conventional MD. For RE simulations, sometimes it helps our understanding to distinguish a "walker" from a "replica". In our terminology, a replica corresponds to a thermodynamic state which stays at a constant temperature while its trajectory in conformational space is frequently interrupted by exchanges. If we follow the view of a walker, however, it leads to a continuous trajectory in conformational space but jumps up and down in temperature space. A very important property of a walker is: in a well-equilibrated RE simulation, every walker should have the same population distribution in temperature space and in conformational space as every other walkers, i.e., all walkers are statistically identical to each other at equilibrium.

When a folding/unfolding transition event occurs for a walker, regardless of what temperature the walker is at during the transition, we call it "a transition event for a walker". Assuming two-state kinetics, the expected number of transition events for a walker (NTE_{walker}) is determined by the harmonic mean of folding and unfolding rate constants at every temperature[51,91]

$$NTE_{walker} = \frac{1}{N}\sum_{i=1}^{N}\frac{\tau}{\frac{1}{k_f(T_i)}+\frac{1}{k_u(T_i)}} \tag{8}$$

where N is the number of replicas, τ is the simulation time for each replica. i.e., when the REMD ensemble equilibrates both in temperature space and conformational space, the total number of folding events from all walkers in the REMD simulation equals the total number of folding events of the same number of uncoupled MD simulations running at the same set of temperatures for the same amount of simulation time. This sets a speed limit for convergence of RE simulations, which is the average of the harmonic mean of the folding and unfolding rate constants at all temperatures. In order to quantify how much RE increases the convergence rate for a target temperature T, we need to define two quantities. $NTE(T)$ is the expected number of folding events at T in a REMD simulation. Assuming that the attempted exchange rate is much faster than the folding/unfolding rate (fast-exchange limit), $NTE(T)$ approaches its upper limit: the number of folding events for a walker: $NTE(T) \to NTE_{walker}$, $as\, k_{RE} \to \infty$, where k_{RE} is the attempted exchange rate. This implies that $NTE(T)$ is the same for all temperatures when the exchange rate is fast and the RE ensemble is at equilibrium. We also define the number of folding events in a MD simulation at T for the same amount of simulation time as $NTE_{MD}(T)$. If $NTE(T) > NTE_{MD}(T)$, RE effectively enhances the sampling at temperature T. The ratio $NTE(T)/NTE_{MD}(T)$ provides an estimate of the amount by which the rate of equilibration at T is increased by REMD compared with MD at the same temperature.

We ran a set of 16 uncoupled MD simulations at the same temperatures as those of the REMD simulation. The starting structure, simulation parameters and total simulation time of the MD simulations are also kept the same as the setup for REMD. With the same definition of folded and unfolded macrostates, the total number of folding events from all walkers ($N \times NTE_{walker}$) is 35 for the REMD simulation and the total number of folding events ($\sum_{i=1}^{N} NTE_{MD}(T_i)$) is 39 for all MD simulations with $N = 16$. They are about equal, as Eq.8 predicts. The distribution of folding events over temperatures of MD simulations are plotted in Fig.9. The number of folding events from the MD simulations are only observed at the highest five temperatures. $NTE(T)$ is about 2.2

per replica calculated from NTE_{walker} in the fast exchange limit. It is indicated as a dashed line in Fig.9. There are zero folding events observed at low temperatures and there are many events at the four highest temperatures in the uncoupled MD simulations . Therefore, it is evident that in the Trp-Cage system, we gain efficiency using REMD for sampling at low temperatures (from 270K to 465K) compared with conventional MD. However, we will lose efficiency if we are trying to use REMD to speed up the sampling at the higher temperatures (from 488 to 566K).

This discussion assumes two-state behavior for the system. For systems that have more complex kinetic behavior (three states or more), the efficiency gain of RE could potentially be higher than the predicted speed limit for convergence in Eq.8. This topic is an ongoing project in our lab.

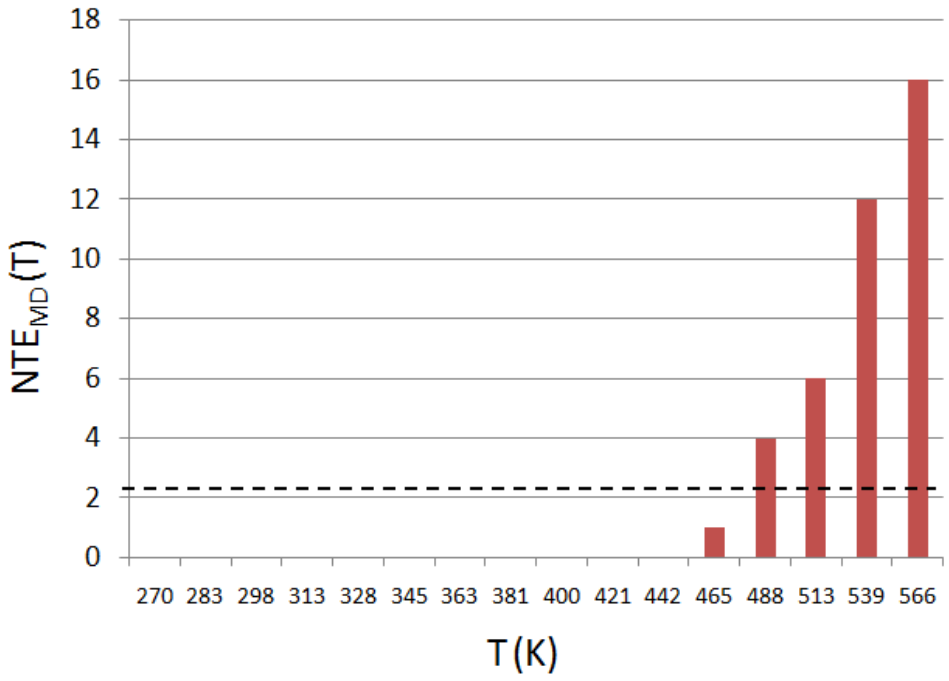
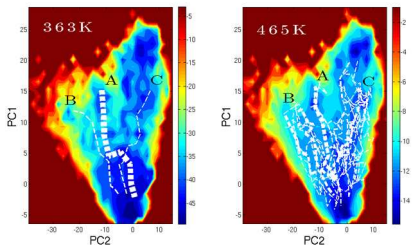


Figure 9: Number of folding events observed at different temperatures for 16 uncoupled MD simulations. The total number of observed folding events from all the temperatures is 39. Folding events are only observed at the highest five temperatures. The dashed line is the expected number of folding events per replica in the REMD simulation $NTE(T)$. For low temperatures (from 270K to 465K), $NTE(T) > NTE_{MD}(T)$, RE effectively enhances the sampling compared with MD.


363K
465K
A
B
C
PC1
PC2