Accepted, J Chem Phys, 2010.

Constrained Proper Sampling of Conformations of Transition State Ensemble during Protein Folding

Ming Lin[1], Jian Zhang[2,3], Hsiao-Mei Lu[2], Rong Chen[4,5], and Jie Liang[2] *

[1] *Wang Yanan Institute for Studies in Economics,*
Xiamen University, Xiamen 361005, China;

[2] *Department of Bioengineering, University of Illinois at Chicago, Chicago, Illinois 60607;*

[3] *National Laboratory of Solid State Microstructure,*
Nanjing University, Nanjing 21093, China

[4] *Department of Statistics, Rutgers University,*
Piscataway, New Jersey 08854-8019.

[5] *Department of Business Statistics and Econometrics,*
Peking University, Beijing 100080, China.

* Corresponding author. Phone: (312)355–1789, fax: (312)996–5921, email: jliang@uic.edu

Abstract

Characterizing the conformations of protein in the transition state ensemble (TSE) is important for studying protein folding. A promising approach pioneered by Vendruscolo *et al*[40] to study TSE is to generate conformations that satisfy all constraints imposed by the experimentally measured ϕ-values that provide information about the native-likeness of the transition states. Faisca *et al*[12] generated conformations of TSE based on the criterion that, starting from a TS conformation, the probabilities of folding and unfolding are about equal through Metropolis Monte Carlo (MC) simulations. In this study, we use the technique of constrained sequential Monte Carlo method (CSMC)[26,44] to generate TSE conformations of acylphosphatase (AcP) of 98 residues that satisfy the ϕ-value constraints, as well as the criterion that each conformation has a folding probability of 0.5 by Metropolis MC simulations. We adopt a two stage process and first generate 5,000 contact maps satisfying the ϕ value constraints. Each contact map is then used to generate 1,000 properly weighted conformations. After clustering similar conformations, we obtain a set of properly weighted samples of 4,185 candidate clusters. Representative conformation of each of these cluster is then selected and 50 runs of Markov chain Monte Carlo (MCMC) simulation are carried using a regrowth move set. We then select a subset of 1,501 conformations that have equal probabilities to fold and to unfold as the set of TSE. These 1,501 samples characterize well the distribution of transition state ensemble conformations of acylphosphatase. Compared with previous studies, our approach can access much wider conformational space and can objectively generate conformations that satisfy the ϕ-value constraints and the criterion of 0.5 folding probability without bias. In contrast to the previous studies, our results show that transition state conformations are very diverse and are far from native-like when measured in cRMSD (Cartesian Root-Mean-Square Deviation): the average cRMSD between TS conformations and the native structure is 9.4Å for this short protein, instead of 6Å reported in previous studies. In addition, we found that the average fraction of native contacts in the TSE is 0.37, with enrichment in native-like β-sheets and a shortage of long range contacts, suggesting such contacts form at a later stage of folding. We further calculate the first passage time of folding of TSE conformations through calculation of physical time associated with the regrowth moves in MCMC simulation through mapping such moves to a Markovian state space model with transition time obtained by Langevin dynamics simulations. Our results indicate

that despite the large structural diversity of the TSE, they are characterized by similar folding time. Our approach is general and can be used to study TSE in other macromolecules.

I. INTRODUCTION

While protein native conformation provides the structural basis of its biological function, it is important to understand how proteins fold to its native state[8,29,37]. Protein folding is a complex process that involves many different molecular and cellular machinery. Protein conformations are inherently heterogeneous and, in many cases, misfolded proteins can cause diseases such as Alzheimer's disease, Parkinson's diseases, and type II diabetes[1]. Characterizing the conformations of transition state ensemble (TSE) of protein folding has been a major focus in protein folding studies[2,15,17,35,38]. Transition state ensemble (TSE) are usually understood to be those conformations around the saddle point of the landscape of protein folding[35]. These conformations have about the same probability to either fold or unfold. Because the transition states are transient in nature, can contain a wide range of conformations, and are often dynamic with significant amount of structural fluctuations, it is challenging to study them with experimental techniques.

An important approach to study TSE is the ϕ-value analysis[16]. By measuring the changes of free energy of activation and free energy of folding upon mutating a residue, this technique provides a measure of the extent of formation of structure relative to denatured and native states of the TSE. Experimental ϕ-value analysis can also provide information on the degree of formation of secondary and tertiary structures[36], backbone-backbone hydrogen-bonding interactions[7], and movement around the transition state in the folding energy landscape[19].

Computational studies have also leaded to important insight on how protein folds. Among these, lattice models and molecular dynamic have been successfully applied to study protein folding and to characterize partially unfolded structures[14]. Klimov and Thirumalai used exhaustive simulations of lattice models with side-chains to study transition state ensemble of two-state folders[22]. Day and Daggett ran multiple molecular dynamic simulations at different temperature and solvent environment to study the folding/unfolding transition state ensemble of chymotrypsin inhibitor 2[6]. Ding *et al* reconstructed the TSE of the src-SH3 protein domain from molecular dynamic simulations[9]. Prompers and Brüschweiler combined molecular dynamics with NMR relaxation spectroscopy to study the dynamics of folded and unfolded proteins[31]. Zagrovic *et al* found that the mean structure averaged over unfolded ensemble of three different folds small proteins are native like[42]. Experimental information, such as NMR residual dipolar couplings, can be used as constraints to select unfolded state structures[27]. Other information from NMR spectroscopy can also be incorporated to define partially folded intermediate states[23,32]. Richter *et al* provided a solution to solve over-fitting and under-fitting problems when calculating ensemble of structures with NMR constraints[33].

To generate explicit conformations of the TSE, Vendruscolo *et al* used information from experimental ϕ-values[28,40]. The ϕ-value at individual residue position is defined as the ratio of stability change to the transition state upon mutation versus stability change of the native folded state upon the same mutation[24,25]. ϕ-values can be measured experimentally and provide rich information about the native-likeness of protein structures in the TSE[13,41]. Following Li and Daggett's work[25], Vendruscolo *et al* defined TSE as the conformations that satisfy

$$\phi_i^{calc} \triangleq \frac{N_i^{TSE}}{N_i^N} \simeq \phi_i^{exp}, \tag{1}$$

for the i-th residue with experimentally measured ϕ-value ϕ_i^{exp}. Here the *calculated ϕ-value* ϕ_i^{calc} of residue-i is defined as the ratio of the number N_i^{TSE} of *native* contacts formed by the residue in the transition state, over the number N_i^N of contacts formed by the residue in the native state. Using Markov chain Monte Carlo method (MCMC) with crank-shaft move, the authors generated a set of TSE conformations based on this model for acylphosphatase (AcP), a protein with 98 residues.

Faisca *et al*[12] used a different approach to identify conformations in the TSE based on the idea that the conformations in TSE have equal probability to either fold or unfold[10]. Starting from a random conformation, independent Monte Carlo (MC) simulations are carried out. If in half of these independent MC runs, the structure folds before unfolds, the initial conformation is identified as a member in the TSE.

In this work, we generate TSE conformations of AcP with the combined constraints of experimental ϕ value as studied by Vendruscolo *et al*[40] and the p_{Fold} criteria[10] as implemented by Faisca *et al*[12]. We use constrained Sequential Monte Carlo to generate candidate conformations that satisfy all ϕ-value constraints. Markov chain Monte Carlo simulations are then carried out to each of the candidate conformations and select only the conformations with folding probability of 0.5. Our main contribution is that, through further development of the technique of constrained Sequential Monte Carlo method first reported in ref[26], we ensure rigorous and efficient sampling of the whole space of TSE under stringent constraints from both ϕ-values and the p_{Fold} model without bias towards native conformations due to inadequate sampling in molecular dynamics simulation, or the unsolved difficulty in assessing adequate mixing when applying Metropolis type of Monte Carlo sampling techniques.

This paper is organized as follows. In Section 2, we described our method to generate conformations in TSE for the protein AcP. Findings and interpretations of the reproduced TSE are reported in Section 3, followed by the Conclusion Section.

II. MODEL AND METHOD

A. Generating candidate conformations of TSE

We first generate a set of candidate conformations of the transition state ensemble of AcP that satisfy the constraints of all experimentally measured ϕ-values at different positions of amino acid residues. Here we follow Vendruscolo *et al*'s model of ϕ-value constraints[40]. Specifically, our goal is to generate a proper set of conformations that are uniformly distributed in the model constrained space

$$\Omega_\phi = \{\boldsymbol{x}_n : |\phi_i^{calc} - \phi_i^{exp}| < 0.15 \quad \text{for all } i \in \mathcal{I}\}, \tag{2}$$

where $\boldsymbol{x}_n = (x_1, \cdots, x_n)$ denotes a conformation of the protein, which has n residues. x_i is the location of i-th residue, ϕ_i^{exp} and ϕ_i^{calc} are the experimentally measured ϕ-value and the *calculated* ϕ-value of the i-th residue, respectively; $\mathcal{I}$ is the set of residues whose ϕ-values have been measured experimentally.

We consider a three-dimensional cubic lattice model, in which residues in conformation $\boldsymbol{x}_n$ are located on the lattice sites with a unit length of 1.3 Å and satisfy the self-avoiding, bond-length, bond-angle, and torsion-angle constraints. It is based on an off-lattice 4-state model, and on average there are 23 candidate positions for placing an additional residue to a partial chain[26]. Two residues are defined to be in contact if the distance between them is less than 8.5 Å . Details of this lattice model and constraints are described in[26,44]. In this lattice model for protein AcP, the conformation that is closest to the native structure in terms of cRMSD has 88% native contacts preserved.

We use the sequential Monte Carlo technique to generate AcP structures. It is a growth-based method that can generate samples $\{(\boldsymbol{x}_n^{(j)}, w^{(j)}), j = 1, \cdots, m\}$ properly weighted with respect to a given target distribution $\pi(\boldsymbol{x}_n)$. The weights are calculated as $w^{(j)} \triangleq w(\boldsymbol{x}_n^{(j)}) = \pi(\boldsymbol{x}_n^{(j)})/q(\boldsymbol{x}_n^{(j)})$, where $q(\boldsymbol{x}_n^{(j)})$ is the probability of generating the sample $\boldsymbol{x}_n^{(j)}$. If this sampling distribution satisfying $q(\boldsymbol{x}_n) > 0$ for all $\boldsymbol{x}_n \in \{\boldsymbol{x}_n \mid \pi(\boldsymbol{x}_n) > 0\}$, any function $h(\boldsymbol{x}_n)$ under the target distribution $\pi(\boldsymbol{x}_n)$ can be estimated by

$$\widehat{\mathbb{E}}_\pi\big(h(\boldsymbol{x}_n)\big) = \frac{\sum_{j=1}^m w^{(j)} h(\boldsymbol{x}_n^{(j)})}{\sum_{j=1}^m w^{(j)}}. \tag{3}$$

In addition, the normalizing constant of the target distribution $\pi(\boldsymbol{x}_n)$ in any set Ω, namely, the partition function in the case when the target distribution is the Boltzmann distribution, can be estimated using

$$\sum_{\boldsymbol{x}_n \in \Omega} \pi(\boldsymbol{x}_n) \approx \frac{1}{m} \sum_{j=1}^{m} w^{(j)} \cdot \mathbb{I}(\boldsymbol{x}_n^{(j)} \in \Omega), \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function: $\mathbb{I}(\cdot) = 1$ if the statement represented by $(\cdot)$ is true, 0 otherwise.

Lin *et al*[26] used a two-stage sequential Monte Carlo method to efficiently generate conformation samples properly weighted with respect to the uniform distribution in Ω_ϕ, that is, $\pi(\boldsymbol{x}_n) \propto \mathbb{I}(\boldsymbol{x}_n \in \Omega_\phi)$. At the first stage, 5,000 contact maps are sampled from the uniform distribution of all contact maps satisfying the ϕ-value constraints. Here each sample is a realization of a $n \times n$ symmetric contact map $\mathcal{C} = \{c_{ij}\}_{n \times n}$, where $c_{ij} = 1$ if residue i and residue j are in contact, and $c_{ij} = 0$ otherwise. At the second stage, for each contact map sample, $1,000$ properly weighted conformational samples satisfying this contact map are generated. For protein AcP, Fig. 1 shows the experimentally measured ϕ-values and the weighted average of the calculated ϕ-values of the generated conformation samples.

To reduce the number of candidate conformations, we cluster similar conformations together. First, we arrange all the conformations in a random order. Starting from an empty set, we add one conformation at a time to the current system of clusters, from the first conformation to the last conformation. For each conformation, it is compared with all the current cluster representatives. If its cRMSD to any previous clusters is larger than a cutoff-value, it is regarded as being a member of a new cluster; otherwise, it is grouped with the nearest cluster. The cutoff value for clustering used in this study is 2 Å . The weight of each cluster is the summation of the weights of all conformations in that cluster, and the representative structure of each cluster is chosen as the conformation with the largest weight in that cluster. For protein AcP, we obtained a total 4,185 clusters. The fraction of native contacts preserved in these clusters is within a small range of 0.15, namely, from 0.26 to 0.41. This is not surprising because of the strong ϕ-value constraints imposed.

B. Identifying conformations in TSE using Markov chain Monte Carlo

a. p_{fold} *estimated by Markov chain Monte Carlo:* In addition to the constraints from measured ϕ-values, we further adopt the p_{fold} model introduced in[5,11] in which the transition state conformation will have about equal probability to fold or unfold. According to[5,11], in a system with two stable states (the folded state and the unfolded state), the folding probability, p_{fold} of any conformation is defined as the probability that it will reach the folded state before reaching the unfolded state. p_{fold} can be regarded as a measure of the kinetic distance between the given conformation and the folded state. It is therefore reasonable to assume that the conformations in the TSE would have $p_{\text{fold}} = 0.5$. Starting from a specific conformation, Faisca *et al* calculates p_{fold} of the conformation by recording the ratio of runs of Markov chain Monte Carlo simulations that reach the folded state before reaching the unfolded state[12]. The conformations of TSE are then obtained by selecting those conformations with $p_{\text{fold}} = 0.5$. We follow this strategy to compute p_{fold} for candidate conformations that satisfy the ϕ-value constraints.

Briefly, we construct a Markov chain $\boldsymbol{z}_n^{(1)}, \boldsymbol{z}_n^{(2)}, \cdots, \boldsymbol{z}_n^{(t)}, \cdots$ for the target equilibrium distribution $\pi(\boldsymbol{z}_n)$ of Boltzmann distribution by the Gō-potential as follows[34]: Starting with $\boldsymbol{z}_n^{(1)} = \boldsymbol{x}_n$, where $\boldsymbol{x}_n$ is one of the candidate conformations; at each step t, a random move selected from a primitive move set is applied to $\boldsymbol{z}_n^{(t-1)}$ to obtain a new conformation $\boldsymbol{z}_n^{\text{new}}$.

$\boldsymbol{z}_n^{\mathrm{new}}$ is accepted as $\boldsymbol{z}_n^{(t)}$ with probability

$$\min\left\{1, \frac{g(\boldsymbol{z}_n^{(t-1)} \mid \boldsymbol{z}_n^{\mathrm{new}})\pi(\boldsymbol{z}_n^{\mathrm{new}})}{g(\boldsymbol{z}_n^{\mathrm{new}} \mid \boldsymbol{z}_n^{(t-1)})\pi(\boldsymbol{z}_n^{(t-1)})}\right\},$$

and let $\boldsymbol{z}_n^{(t)} = \boldsymbol{z}_n^{(t-1)}$ otherwise. Here $g(\boldsymbol{z}_n^{\mathrm{new}} \mid \boldsymbol{z}_n^{\mathrm{old}})$ is the probability of moving from the current conformation $\boldsymbol{z}_n^{\mathrm{old}}$ to the new conformation $\boldsymbol{z}_n^{\mathrm{new}}$.

b. Re-growth move set: We use the primitive move set developed by Zhang *et al*[43] in this study. The primitive move is to randomly remove a fragment of the current conformation $\boldsymbol{z}_n^{(t-1)}$, and regenerate the removed fragment to obtain a new conformation $\boldsymbol{z}_n^{\mathrm{new}}$. The fragment is regenerated using sequential Monte Carlo under the constraint that the two ends of the fragment are fixed. If the removed fragment is at the tail of the conformation, only one end is fixed. The starting position of the fragment to be replaced is uniformly distributed along the full chain, and the fragment length is uniformly distributed between 5 and 12.

c. Folded and unfolded state: We assess whether a Markov chain $\{\boldsymbol{z}_n^{(t)}\}$ at time t has reached the folded state or unfolded state by criteria based on the number of native contacts preserved in the conformation $\boldsymbol{z}_n^{(t)}$. We set two thresholds N_{fold} and N_{unfold} for the number of native contacts in a conformation. If the number of native contacts preserved in $\boldsymbol{z}_n^{(t)}$ is larger than N_{fold}, the conformation is considered to be folded. If it is less than N_{unfold}, the conformation is considered to be unfolded.

The values of N_{fold} and N_{unfold} are determined as follows. For N_{fold}, we sample uniformly from the set of near native conformations (NNS) Ω_{NNS}. Here we follow[44] and define the set of NNS as those within 3 Å in cRMSD from the native structure. N_{fold} is defined as the threshold value of number of native contacts, such that only 5% of the conformations in Ω_{NNS} have less than N_{fold} native contacts. For N_{unfold}, we sample uniformly from the set of denatured conformations Ω_{D}, defined as the set of conformations with > 10 Å in cRMSD from the native structure. N_{unfold} is defined as the threshold value of number of native contacts, such that only 5% of the conformations in Ω_{unfold} have more than N_{unfold} native contacts. Since the majority of the conformations in the set of all possible conformations have cRMSD to the native structure > 12 Å, the choice of the value of 10 Å for deriving N_{unfold} is not critical.

We use the sequential Monte Carlo method described in[44] to generate sets of proper weighted conformations in both Ω_{NNS} and Ω_{D}. Fig. 2(a) shows the values of the two thresholds for N_{fold} and N_{unfold} for protein AcP. They satisfy $N_{\text{fold}}/N = 0.65$ and $N_{\text{unfold}}/N = 0.15$, where N is the number of contacts in the native structure.

d. The energy function in Markov chain Monte Carlo: In the Markov chain Monte Carlo runs, the energy function we use is the Gō-potential[18]

$$H(\boldsymbol{x}_n) = \sum_{i>j+3} U(x_i, x_j),$$

where $U(x_i, x_j) = -1$ only if residue i and j are in contact, namely, $|x_i - x_j| < 8.5\,\text{Å}$, in both conformation $\boldsymbol{x}_n$ and the native structure. The equilibrium distribution of the Markov chain $\{\boldsymbol{z}_n^{(t)}\}$ is $\pi(\boldsymbol{x}_n) \propto \exp\{-H(\boldsymbol{x}_n)/\tau\}$, where τ is the temperature parameter. Here we use s slightly different definitions of the set of NNS and the set of denatured conformations, based on N_{fold} and N_{unfold}:

$$\Omega^*_{\text{NNS}} \overset{\triangle}{=} \{\boldsymbol{x}_n \,|\, \text{number of native contacts preserved in } \boldsymbol{x}_n \text{ is larger than } N_{\text{fold}}\},$$

$$\Omega^*_{\text{denature}} \overset{\triangle}{=} \{\boldsymbol{x}_n \,|\, \text{number of native contacts preserved in } \boldsymbol{x}_n \text{ is less than } N_{\text{unfold}}\}.$$

Following Ref.[12], the folding temperature is selected so that the folded structures and the denatured structures have equal probabilities in the equilibrium distribution. That is,

$$\sum_{\boldsymbol{x}_n \in \Omega^*_{\text{NNS}}} \exp\{-H(\boldsymbol{x}_n)/\tau\} \simeq \sum_{\boldsymbol{x}_n \in \Omega^*_{\text{denature}}} \exp\{-H(\boldsymbol{x}_n)/\tau\}. \tag{5}$$

For a specific given temperature τ, We again use the sequential Monte Carlo technique to estimate the values of both sides of Eq. (5)[44]. For protein AcP, the temperature is set to $\tau = 1.654$, which makes both sides of Eq. (5) equal.

We carry out 50 Markov chain Monte Carlo runs for each of the 4,185 conformations that satisfy the ϕ-value constraints. We then test the null hypothesis $p_{\text{fold}} = 0.5$. This null hypothesis is rejected if the statistical p-value of the number of runs that lead to the folded state is less than 5%. Or equivalently, if the number of runs that fold is less than 17 or larger than 33 among the 50 independent runs starting from $\boldsymbol{x}_n$, we reject the null hypothesis that $p_{\text{fold}} = 0.5$, and this conformation $\boldsymbol{x}_n$ is not included in the TSE. Otherwise, $\boldsymbol{x}_n$ is included in the TSE. Fig. 2(b) shows the number of runs that lead to the folded state for 4,185 conformations. A total 1,501 conformations are included in the TSE.

III. RESULTS

In this section, we study the physical properties of conformations that form the TSE of the protein AcP using the aforementioned procedure. The generated samples representing the TSE of AcP consist of 1,501 clusters of conformations, each conformation is associated with a properly calculated weight with respect to the Boltzmann distribution $\pi(\boldsymbol{x}_n) \propto \exp\{-H(\boldsymbol{x}_n)/\tau\}$. As the weight obtained in Section II A is with respect to the uniform distribution in the constrained space Ω_ϕ, they have been adjusted by a multiplication factor $\exp\{-H(\boldsymbol{x}_n)/\tau\}$.

A. TSE can be far away from the native state

We plot the distribution of cRMSD between TS conformations and the native state and the distribution of the fraction of native contacts preserved in TS conformations in Fig. 3(a) and (b), respectively. The un-weighted transition state ensemble has a large variation, with the cRMSD ranging from 6 Å to 14 Å and the fraction of native contacts preserved varying from 0.28 to 0.38. In contrast, the transition state ensemble weighted with respect to the Boltzmann distribution is much more homologous - the majority of conformations have a cRMSD of 9.4 Å and fraction of native contacts preserved of 0.37. The difference between un-weighted and weighted TSE demonstrates that TS conformations are structurally diverse. Although the weighted TSE is much more homologous than the un-weighted one, the average cRMSD remains to be large, compared with the value of 6 Å reported in a previous work[40].

This difference is likely due to the fact that our method can access much wider conformational space in severely constrained space. As a result, TS conformations that are far away from the native state are successfully identified, and are represented proportionately with correct importance weights that adjusts the sampling bias for using a sampling distribution that is different from the target distribution. The weight ensures that it neither exaggerates nor underestimates the importance of these conformations in the TSE that are far away from the native state. In fact, the characterizations of TSE are accurate for conformations of any nature, including those that are close to the native state.

To compare TS conformations with other conformations satisfying ϕ-value constraints, we divide the 4,185 candidate conformations generated into three groups:

$$\Omega_{\text{TSE}} = \{\boldsymbol{x}_n | \boldsymbol{x}_n \text{ satisfies the } \phi\text{-value constraints and with } p_{\text{fold}} = 0.5\},$$
$$\Omega_{\text{DS}} = \{\boldsymbol{x}_n | \boldsymbol{x}_n \text{ satisfies the } \phi\text{-value constraints and with } p_{\text{fold}} < 0.5\},$$
$$\Omega_{\text{NS}} = \{\boldsymbol{x}_n | \boldsymbol{x}_n \text{ satisfies the } \phi\text{-value constraints and with } p_{\text{fold}} > 0.5\}.$$

That is, if the number of folded runs among the 50 independent Markov chain Monte Carlo simulations starting from $\boldsymbol{x}_n$ is between 17 and 33, the conformation $\boldsymbol{x}_n$ is considered to be in set of transition state ensemble Ω_{TSE}. If the number of folded runs is less than 17, $\boldsymbol{x}_n$ is considered to be in set Ω_{DS} of *denatured side (DS).* If the number of folded runs is larger than 33, $\boldsymbol{x}_n$ is considered to be in set Ω_{NS} of *native side (NS).*

We plot the distribution of cRMSDs between the conformations in these three sets and the native conformation in Fig. 3(a),(c),(e), and the distributions of the fraction of native contacts preserved in these three sets in Fig. 3(b),(d),(f).

Although it appears that many conformations with lower RMSD have small weights as the weighted mean cRMSD is larger (Fig 3e), and there are low energy conformations with large cRMSD that dominate in mean cRMSD calculation, we cannot conclude that in general conformations with higher cRMSD have lower energy. The conformations generated are from a strongly constrained region with both ϕ-value and folding rate constraints imposed. As a result, energy of conformations in this set are not significantly correlated with cRMSD. Fig. 4 shows the plot of energy and cRMSD of the TSE conformations to the native state. There is little correlation between energy and cRMSD of TSE. The estimated correlation coefficient is -0.035, with a p-value of 0.173 for a two-sided t-test of zero correlation.

It is not surprising to see that the conformations in Ω_{DS} have larger cRMSD to the native structure and less native contacts preserved compared to the conformations in Ω_{TSE}. Similarly as expected, we find that conformations in Ω_{NS} have smaller cRMSD to the native structure and contain more native contacts than Ω_{TSE}.

It is informative to examine possible residual secondary structures in the transition state ensemble. AcP protein contains the following secondary structures: β_1 (residues 7-13), α_1 (residues 22-33), β_2 (residues 36-42), β_3 (residues 46-53), α_2 (residues 55-66), β_4 (residues 77-85), and β_5 (residues 93-97).

Fig. 5 shows the distribution of cRMSDs between fragments of secondary structures in the weighted TSE and in the native conformation. We find that although in general that the native secondary structures are not well-preserved in the TSE, fragments of native β-sheets are more enriched in the TSE compared to α-helices. This is consistent with previous study[40].

It has been suggested that the topology of the transition state of AcP is defined by the relative positions of just three "key" residues Y11, P54 and F94[40]. We have carried out additional study using only ϕ-values at these three key residues as constraints. We find that ϕ-values of the other residues can be largely recovered from conformations generated using constraints at the three key residues alone (Fig. 6 (a)). The correlation coefficients between the calculated ϕ-values of all residues recovered using constraints at the three residue and at 24 residues is 0.79. However, the ensemble of conformations generated have overall much larger cRMSD to the native conformations when only three constraints are used (Fig. 6 (b)).

B. Correlation between point-wise distances and ϕ-values

We define the *point-wise* distance of residue i between a conformation and the native conformation as the Euclidean distance between the locations of residue i after optimal rigid superposition of these two conformations. The average point-wise distance of each residue between the weighted TSE and the native state conformations is shown in Fig. 7.

For the 24 residues in AcP with experimentally measured ϕ-values, the correlation between the ϕ-values and the corresponding point-wise distance is -0.574, with a p-value$=0.0017$ for testing zero correlation by a one-sided t-test. The correlation between the calculated ϕ-values of all residues and the corresponding point-wise distances is -0.502, with a p-value of 6.93×10^{-8}. These observations can be rationalized by the physical models of the ϕ-values. If ϕ-value is large, the structure of TSE around the residue is close to the native state, and thus the corresponding point-wise distance is small with many physical contact constraints reflected by the high ϕ-values. If the ϕ-value is small, the structure of TSE around the residue is disrupted, and the corresponding point-wise distance is therefore large.

C. Contact order of TSE

Contact order has been widely used to study the correlation of protein native structures and protein folding rate[30,45]. It is defined as the average residue separation of the contact. We examine the distribution of all native contacts preserved in the weighted TSE at different residue separations in Fig. 8(a). For comparison, the distribution of residue separation for the native conformation in also shown in Fig. 8(b).

We find that the average contact order of native contacts preserved in the weighted TSE is 33.2, while the contact order for the native state is 37.3. Our result shows that there are less long range contacts in the TSE. That is, long range native contacts often occur after protein chains departed from the transition state.

Paci *et al.* provided a detailed study of contact order of TSE for 10 proteins[28]. They added an energy term based on RMSD in ϕ-value to the energy function of molecular mechanics. The contact order of TSE reported here is somewhat different. This is likely due to the difference in the potential function used. Detailed information on how contact order of TSE is related to protein folding rate can be found in[28]. A study based on a modified concept called geometric contact showed that both two-state and multi-state protein folding rate are well correlated to the native state topology[45]. A detailed theoretical study we have carried out on enumerated 2D HP sequences suggests that the folding rate of model proteins of the same native state can differ by 1,000, and the observed correlation of folding rate and native state topology in real proteins may be a consequence of evolutionary selection[20].

D. The first passage time

We now we estimate the first passage time (FPT), which is defined as the average of time required for a conformation in the transition state to fold into its native state. Because the number of Markov moves required for a conformation to fold depends on the specific details of the move set, it usually does not reflect the true physical time required for folding. To arrive at some estimations of the time required for a transition state conformation to fold, we use langevin dynamics simulation to estimate the true physical time that each Markov move takes.

Given the number of residues ($L = 5, \cdots, 12$) in the regrown fragment and the end-to-end distance r of the fragment-ends, we perform MD simulations to estimate the traveling time between different fragment configurations that have the same number of residues L and the same end-to-end distance r. Here we discretize r into bins of intervals between $r = 1.5\,\text{Å}, 2.0\,\text{Å}, 2.5\,\text{Å}, \cdots$, according to the end-to-end distance.

a. Simulation of physical movement of fragment: For a fragment $\boldsymbol{x}$ of length L and end-to-end distance r, we run Langevin dynamics simulations to sample its conformations and calculate the transition time between different conformational clusters. That is, we aim to provide physically relevant time scale for each elementary Monte Carlo move that transform the conformation of a fragment. Since our goal is to assess the physical time of the movement or diffusion of a fragment, we fix its two-ends and measure the time required to transform the conformation of the fragment from $\boldsymbol{x}_L(t_1)$ to $\boldsymbol{x}_L(t_2)$. Here we use a simplified model, in which the residues in the fragment are treated as connected beads, and they are allowed to move freely in the space subjected to the constraints imposed by other residue beads in the fragment through several types of interactions, including the bond interaction, angle interaction, and van der waals interaction. The motion of the system is simulated using Langevin dynamics, where the equation governing the motion of all residues in the fragment is[21]:

$$\frac{d^2\boldsymbol{x}(t)}{dt^2} = -\gamma\frac{d\boldsymbol{x}}{dt} + f(\boldsymbol{x}, t) + \alpha\boldsymbol{\epsilon}(t), \tag{6}$$

where $\boldsymbol{x}(t)$ is the position vector of the residues at time t, γ is the friction constant, $f(\boldsymbol{x}, t)$ is the conformational force per unit mass, and α is a constant defined as $\alpha = (2\gamma T/m)^{\frac{1}{2}}$, in which T is the temperature, m is the mass, and $\boldsymbol{\epsilon}(t)$ is the Gaussian random force at time t, such that the autocorrelation function $< \boldsymbol{\epsilon}(t), \boldsymbol{\epsilon}(t') >= \delta(t - t')$, where $\delta(t)$ is the delta function. Here we have $\gamma = 0.05\tau^{-1}$, with $\tau = \sqrt{m \cdot l \cdot l/e}$ being the time unit of the simulation. $m = 1$ is the mass unit, $l = 3.8$Å is the length unit, and $e = 1$ is the energy unit. Veitshans *et al.* provided a discussion on the choice of the value of the friction constant γ[39].

For each combination of L and r, we start from a chain in an extended initial conformation and an initial velocity vector in Gaussian form, in which each of the $3L$ vector component is sampled from the Gaussian distribution $\mathcal{N}(0, 1)$, which is then scaled by a factor of $\sqrt{T}$. The simulation is run for 10^9 time steps, where each time step is set to $0.005\sqrt{ma^2/\epsilon_0}$, with mass $m = 1$, length scale $a = 3.8$ Å, and the reference energy scale $\epsilon_0 = 1$.

The first 2×10^8 steps is treated as the burning-in period and the generated fragment conformations are discarded. The fragment conformations beyond the burning period are clustered as follows. Each time a conformation is sampled, it is compared with all the cluster representatives generated in previous steps. If its distance is more than a cut-off threshold from all of the representatives, it is considered as the representative of a new cluster; otherwise, it is grouped to its nearest cluster. Here the distance between two fragments $\boldsymbol{x}_L(t_1)$ at time t_1 and $\boldsymbol{x}_L(t_2)$ at time t_2 is calculated as:

$$d\left(\boldsymbol{x}_L(t_1),\, \boldsymbol{x}_L(t_2)\right) = \left[\frac{1}{L-2}\sum_{l=1}^{L} |x_l(t_1) - x_l(t_2)|^2\right]^{1/2} \tag{7}$$

in which $|x_l(t_1) - x_l(t_2)|$ is the Euclidean distance between the two position vectors $x_l(t_1)$ and $x_l(t_2)$ of the l-th residue. The cutoff used in the clustering is 5 Å.

b. Markovian assumption and the estimation of traveling time: Suppose S clusters are obtained. We treat each cluster as a state, and use state i to denote the i-th cluster. The representative structure of state i is denoted as $\boldsymbol{y}_L^i$. Let I be the total number of time steps of the trajectory beyond the burning-in period, I_i be the observed number of state i and I_{ij} be the observed number of times that state i is immediately followed by state j in the next time step. We define $\widehat{p}_i = I_i/I$, which represents the probability of the fragment to be in state i, and $\widehat{p}_{ij} = I_{ij}/I_i$, which represents the transition probability from state i to j.

The average duration ξ_i that the state sequence of the simulation trajectory $\{\boldsymbol{x}_L(t)\}$ stays in state i can be estimated as:

$$\xi_i = \sum_{k=0}^{\infty} k\widehat{p}_{ii}^{k}(1 - \widehat{p}_{ii}) = \widehat{p}_{ii}/(1 - \widehat{p}_{ii}), \tag{8}$$

where $1 - \widehat{p}_{ii}$ represents the probability of the fragment moves away from state i.

To estimate the average time ξ_{ji} that the state sequence enters state j ($j \neq i$), then travels from state j to state i, we analyze the time trajectory. If the state sequence $\{\boldsymbol{x}_L(t)\}$ leaves state i at step t_0 then re-enters state i at step t_1, we record the first time that $\{\boldsymbol{x}_L(t)\}$ enters state j after t_0 but before t_1 as $t(j)$. The traveling time $\widetilde{\xi}_{ji}$ is then recorded as $t_1 - t(j)$. As many $\widetilde{\xi}_{ji}$ can be recorded from one simulation trajectory, we take its average value as the travel time ξ_{ji}. An illustration of counting $\widetilde{\xi}_{ji}$ is shown in Fig 9(a).

after clustering is a Markov chain, we can alternatively calculate ξ_{ji}, $j \neq i$ for each state i by solving the linear equations

$$1 + \sum_{k \neq i} \widehat{p}_{jk} \xi_{ki} = \xi_{ji}, \quad \text{with} \quad j = 1, \cdots, i-1, i+1, \cdots, S.$$

Fig. 10 shows the frequency of different states in the MD simulation and the comparison between the transition time calculated through counting the simulated MD sequence, namely, the *counted* traveling time, and that through solving the linear equations, namely, the *calculated* traveling time. ¿From Fig. 10 we observe that the ratio of the counted and the calculated traveling times is close to one, except for those states with very few observations. The generally good agreement between these two approaches suggests that a Markovian state model is reasonable for the majority of state transitions.

c. Physical time for the regrowth moves: After obtaining the traveling time ξ_{ij}, the time each Markov move takes is estimated as follows. Suppose the Markov chain moves from the current fragment $\boldsymbol{x}_L^{\text{old}}$ to a new fragment $\boldsymbol{x}_L^{\text{new}}$. First, we assign $\boldsymbol{x}_L^{\text{old}}$ to the state i, whose representative structure $\boldsymbol{y}_L^{(i)}$ is the closest to $\boldsymbol{x}_L^{\text{old}}$ in terms of cRMSD. If the proposed move is rejected, this move takes time ξ_i. If the move is accepted, we assign the new fragment $\boldsymbol{x}_L^{\text{new}}$ to a new state j, in which $\boldsymbol{y}_L^{(j)} - \boldsymbol{y}_L^{(i)}$ is the closest to $\boldsymbol{x}_L^{\text{new}} - \boldsymbol{x}_L^{\text{old}}$ in terms of Euclidean distance. This successful move takes time $\xi_{ij} - \xi_i$, as we assume that the fragment will stay in state i on average ξ_i time before it moves to state j.

d. Conformations in TSE have diverse structures but share similar characteristic folding time: Fig. 9(b) plots the average FPT for each conformation in the TSE against the fraction of native contacts in this conformation. The correlation coefficients between the folding time and the fraction of native contacts preserved for the conformations in TSE is -0.068 (p-value of testing zero correlation is 0.0041). Hence the folding time and the nativeness of the conformation are not strongly correlated.

For unweighted TSE, the standard deviation of the average FPT between different conformations is 1.26×10^8 unit of time. For comparison, we computed the standard deviation of FPT for each conformation in different Markov chain Monte Carlo runs, and the average is 3.77×10^8 unit of time. For weighted TSE, these values are 1.98×10^8 and 2.59×10^8, respectively. We can see that the variation of the average FPT for different TS conformations is small. In fact, it is comparable with the variation of FPT in different Monte Carlo runs starting with the same TS conformation. This result shows that, for the protein AcP, although the conformations in TSE are structurally diverse and far away from the native state, they have very similar physical folding time. One possible reason is that, as demonstrated by Fig. 3, all conformations in TSE have relative high energy, therefore these conformations may quickly fold to conformations that have low energy. As a result, these structurally diverse conformations demonstrate similar folding time. Fig. 11(a) plots the average first passage time of TS conformations in different intervals of cRMSD distance to the native structure. It shows that for TS conformations, the average first passage time does not change much as the cRMSD distance to the native structure increases.

We compare the first passage time for the conformations in Ω_{DS} and Ω_{NS} with the TS conformations. Note these groups of conformations are defined by whether they will first fold or unfold by the p_{fold} criterion, without considerations of their kinetic behavior. The distributions of the first passage time for the conformations in these three sets are plotted in Fig. 11(b), (c), and (d). It is not surprising to observe that compared with the conformations in TSE, the average folding time of the conformations in Ω_{DS} is longer, and the average folding time of the conformations in Ω_{NS} is shorter.

IV. DISCUSSION AND CONCLUSIONS

In this study, we have further developed the constrained sequential Monte Carlo method for sampling conformations of transition state ensemble of protein folding. Our approach can generate rigorously unbiased samples for a specified target distribution satisfying experimentally measured parameters such as ϕ values, and can access a much wider space of conformations compared to other methods, and hence lead to generation of more diverse conformations. When combined with Markov chain Monte Carlo with physically mapped transition time, we can generate explicitly conformations of the TSE satisfying both the ϕ value measurement and the p_{fold} criterion.

Our method was applied to study the TSE of the protein acylphosphatase, which has 98 residues. We found that the transition state conformations are diverse, and can be far away from the native state. Although in general native secondary structures are not well-conserved, fragments of native beta sheets are more enriched in the TSE than alpha helices. In addition, we found that long range native contacts are formed only after the formation of TSE. Despite the significant diversity in structures, all TS conformations have similar folding time.

As demonstrated by Cavalli *et al*[3], there is a strong tendency that the outcome of p_{fold} analysis depends on the potential function. It is expected that the Gö potential may introduce a strong bias towards the native state. This would enable structures far away from the native state to have $p_{\text{fold}} = 0.5$. However, the finding of more heterogeneous nature of TSE in this study is most likely due to the improved simulation method employed, and possibly not so much as a consequence of the Gö potential used. The Gö approach is used in the study of Vendruscolo *et al.*[40], in which the potential is a function of RMSD deviation of ϕ-value from the native state. The current results are obtained under comparable settings with these prior studies. Hence the more heterogeneous nature of the TSE is indeed a novel finding of this study.

A challenge in constrained sequential Monte Carlo sampling is to identify an efficient approximating trial distribution $q(x_n)$ in a high-dimensional and strongly constrained space. To reduce the estimate variance, we use carefully designed growth potential described in[26] to generate conformations. In addition, a large sample size $(5,000,000)$ is used to improve accuracy in estimation.

Since our goal is to access the wide conformational space that satisfies all ϕ-value constraints, we use the uniform distribution in the constrained space as the target distribution $\pi(\boldsymbol{x})$. The growth potential is also designed for the uniform target distribution, and is well suited for this purpose. Nevertheless, when we reweight the conformations by Boltzmann factor under the Gō-potential, weights of generated samples become skewed. As Gō-potential models themselves are artificial constructs, it is appropriate to study the natural underlying shapes of the transition state ensemble, which follow the uniform distribution. With this goal in mind, the TSE conformations are generated uniformly from the space with constrained ϕ-values.

In our clustering method, the choices of the representative structures are important because the distance of a conformation to the representatives is used for classification. Although the clustering results may depend on the order in which the conformations are generated, the representative structures are always chosen as those with the largest weights in clusters, regardless of the ordering of the conformations. In addition, by carefully choosing the criterion of cluster distance, conformations are all well separated. We therefore expect that our clustering method is not overly sensitive to differences in the ordering of the conformations. To confirm it, we carried out the following study. We first order the conformations by their weights, then perform clustering sequentially from the largest weight conformation to the smallest weight conformation. This approach resulted in 3,897 clusters, compared to the 4,185 clusters obtained with random ordering. Fig. 12 reports the unweighted distributions of cRMSD values and fractions of native contacts preserved for clusters obtained under both ordering. It is seen that the two ordering produces very similar results.

V. ACKNOWLEDGEMENT

We thank Dr. Eugene Shahknovich for suggesting the p_{fold} model, Drs. Ulrich Hansmann and Hagai Meirovich for helpful discussions. and two anonymous referees for their constructive comments. This work is supported by NIH grants GM079804, GM081682, NSF grant DBI-0646035, and DMS-0800257, DMS-0800183, DMS-0915139 and DMS-0905763.

[1] M. Bartolini and V. Andrisano. Strategies for the inhibition of protein aggregation in human diseases. *Chembiochem*, 11(8):1018–35, Apr 2010.

[2] N. Calosci, C.N. Chi, B. Richter, C. Camilloni, A. Engstrom, L. Eklund, C. Travaglini-Allocatelli, S. Gianni, M. Vendruscolo, and P. Jemth. Comparison of successive transition states for folding reveals alternative early folding pathways of two homologous proteins. *Proc Natl Acad Sci U S A*, 105(49):19241–19246, Dec 2008.

[3] A. Cavalli, M. Vendruscolo, and E. Paci. Comparison of sequence-based and structure-based energy functions for the reversible folding of a peptide. *Biophys J*, 88:3158–66, 2005.

[4] F. Chiti, N. Taddei, P. M. White, M. Bucciantini, F. Magherini, M. Stefani, and C. M. Dobson. Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat. Struct. Biol.*, 6:1005–1009, 1999.

[5] S. S. Cho, Y. Levy, and P. G. Wolynes. P versus q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proc. Natl Acad. Sci.*, 103:586–591, 2006.

[6] R. Day and V. Daggett. Sensitivity of the folding/unfolding transition state ensemble of chymotrypsin inhibitor 2 to changes in temperature and solvent. *Protein Science*, 14:1242–1252, 2005.

[7] S. Deechongkit, H. Nguyen, E.T. Powers, P.E. Dawson, M. Gruebele, and J.W. Kelly. Context-dependent contributions of backbone hydrogen bonding to beta-sheet folding energetics. *Nature*, 430(6995):101–105, Jul 2004.

[8] K.A. Dill, S.B. Ozkan, M.S. Shell, and T.R. Weikl. The protein folding problem. *Annu Rev Biophys*, 37:289–316, 2008.

[9] F. Ding, W. H. Guo, N. V. Dokholyan, E. I. Shakhnovich, and J. E. Shea. Reconstruction of the src-sh3 protein domain transition state ensemble using multiscale molecular dynamics simulations. *J. Mol. Biol.*, 350:1035–1050, 2005.

[10] R. Du, V.S. Pande, A. Y. Grosberg, T. Tanaka, and E.S. Shakhnovich. On the transition coordinate for protein folding. *Journal of Chemical Physics*, 108(1):334–350, 1998.

[11] R. Du, V.S. Pande, A.Y. Grosberg, T. Tanaka, and E. I. Shakhnovich. On the transition coordinate for protein folding. *J. Chem. Phys.*, 108:334–350, 1998.

[12] P. F. N. Faisca, R. D. M. Travasso, R. C. Ball, and E. I. Shakhnovich. Identifying critial residues in protein folding: Insights from ϕ-value and p_{fold} analysis. *J. Chem. Phys.*, 129:095108, 2008.

[13] A. R. Fersht, R. J. Leatherbarrow, and T. N. Wells. Structure-activity relationships in engineered proteins. analysis of use of binding energy by linear free energy relationships. *Biochemistry*, 26:6030–6038, 1987.

[14] A.R. Fersht and V. Daggett. Protein folding and unfolding at atomic resolution. *Cell*, 108:573–582, 2002.

[15] A.R. Fersht, L.S. Itzhaki, N.F. elMasry, J.M. Matthews, and D.E. Otzen. Single versus parallel pathways of protein folding and fractional formation of structure in the transition state. *Proc*

Natl Acad Sci U S A, 91(22):10426–10429, Oct 1994.

[16] A.R. Fersht and S. Sato. Phi-value analysis and the nature of protein-folding transition states. *Proc Natl Acad Sci U S A*, 101(21):7976–7981, May 2004.

[17] C.D. Geierhaas, X. Salvatella, J. Clarke, and M. Vendruscolo. Characterisation of transition state structures for protein folding using 'high', 'medium' and 'low' Phi-values. *Protein Eng Des Sel*, 21(3):215–222, Mar 2008.

[18] N. Go and H. Taketomi. Respective roles of short- and long-range interactions in protein folding. *Proc. Natl. Acad. Sci. U.S.A.*, 75:559–563, 1978.

[19] L. Hedberg and M. Oliveberg. Scattered hammond plots reveal second level of site-specific information in protein folding: phi' (beta++). *Proc Natl Acad Sci U S A*, 101(20):7606–7611, May 2004.

[20] S. Kachalo, H.M. Lu, and J. Liang. Protein folding dynamics via quantification of kinematic energy landscape. *Phys Rev Lett*, 96(5):058106, 2006.

[21] H. Kaya and H. S. Chan. Solvation effects and driving forces for protein thermodynamic and kinetic cooperativity: How adequate is native-centric topological modeling? *J. Mol. Biol.*, 326:911C931, 2003.

[22] D. K. Klimov and D. Thirumalai. Multiple protein folding nuclei and the transition state ensemble in two state proteins. *Proteins Struct. Funct. Gen.*, 43:465–475, 2001.

[23] D. M. Korzhnev, X. Salvatella, M. Vendruscolo, A. A. Di Nardo, A. R. Davidson, C. M. Dobson, and L. E. Kay. Low-populated folding intermediates of fyn sh3 characterized by relaxation dispersion nmr. *Nature*, 430:586–590, 2004.

[24] T. Lazaridis and M. Karplus. "New View" of protein folding reconciled with the old through multiple unfolding simulations. *Science*, 278:1928–1931, 1997.

[25] A. Li and V. Daggett. Characterization of the transition state of protein unfolding by use of molecular dynamics: Chymotrypsin inhibitor. *Proc. Natl Acad. Sci. USA*, 91:10430–10434, 1994.

[26] M. Lin, H. Lu, R. Chen, and J. Liang. Generating properly weighted ensemble of conformations of proteins from sparse or indirect distance constraints. *J. Chem. Phys.*, 129:094101, 2008.

[27] G. A. Mueller, W. Y. Choy, D. Yang, J. D. Forman-Kay, R. A. Venters, and L. E. Kay. Global folds of proteins with low densities of noes using residual dipolar couplings: application to the

370-residue maltodextrin-binding protein. *J. Mol. Biol.*, 300:197–212, 2000.

[28] E. Paci, K. Lindorff-Larsen, C.M. Dobson, M. Karplus, and M. Vendruscolo. Transition state contact orders correlate with protein folding rates. *J. Mol. Biol.*, 352:495–500, 2005.

[29] V.S. Pande, A.Yu. Grosberg, T. Tanaka, and D.S. Rokhsar. Pathways for protein folding: is a new view needed? *Curr Opin Struct Biol*, 9(1):68–79, Feb 1998.

[30] K.W. Plaxco, K.T. Simons, I. Ruczinski, and D. Baker. Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics. *Biochemistry*, 39:11177–83, 2000.

[31] J. J. Prompers and R. Brüschweiler. General framework for studying the dynamics of folded and nonfolded proteins by nmr relaxation spectroscopy and md simulation. *J. Am. Chem. Soc.*, 124:4522–4534, 2002.

[32] T. L. Religa, J. S. Markson, U. Mayor, S. M. Freund, and A. R. Fersht. Solution structure of a protein denatured state and folding intermediate. *Nature*, 437:1053–1056, 2005.

[33] B. Richter, J. Gsponer, P. Varnai, X. Salvatella, and M. Vendruscolo. The mumo (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *J. Biomol. NMR*, 37:117–135, 2007.

[34] R. Robert and G. Casella. *Monte Carlo statistical methods.* Springer, New York, 1999.

[35] C.A. Royer. The nature of the transition state ensemble and the mechanisms of protein folding: a review. *Arch Biochem Biophys*, 469(1):34–45, Jan 2008.

[36] S. Sato, T.L. Religa, V. Daggett, and A.R. Fersht. Testing protein-folding simulations by experiment: B domain of protein a. *Proc Natl Acad Sci U S A*, 101(18):6952–6956, May 2004.

[37] E. Shakhnovich. Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. *Chem Rev*, 106(5):1559–1588, May 2006.

[38] P. Varnai, C.M. Dobson, and M. Vendruscolo. Determination of the transition state ensemble for the folding of ubiquitin from a combination of phi and psi analyses. *J Mol Biol*, 377(2):575–588, Mar 2008.

[39] T. Veitshans, D. Klimov, and D. Thirumalai. Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Folding and design*, 2:1–22, 1996.

[40] M. Vendruscolo, E. Paci, C.M. Dobson, and M. Karplus. Three key residues form a critical contact network in a protein folding transition state. *Nature*, 409:641–645, 2001.

[41] G. Winter, A. R. Fersht, A. J. Wilkinson, M. Zoller, and M. Smith. Redesigning enzyme structure by site-directed mutagenesis : Tyrosyl tRNA ATP binding. *Nature*, 299:756–758, 1982.

[42] B. Zagrovic, C. Snow, S. Khaliq, M. Shirts, and V. S. Pande. Native-like mean structure in the unfolded ensemble of small proteins. *J. Mol. Biol.*, 323:153–164, 2002.

[43] J. Zhang, S. C. Kou, and J. S. Liu. Polymer strucutre optimization and simulation via a fragment re-growth monte carlo. *J. Chem. Phys.*, 126:225101, 2007.

[44] J. Zhang, M. Lin, R. Chen, J. Liang, and J. S. Liu. Monte Carlo sampling of near-native structures of proteins and applications. *Proteins*, 66:61–68, 2007.

[45] O. Zheng and J. Liang. Predicting protein folding rates from geometric contact and amino acid sequence. *Protein Science*, 17:1256–1263, 2008.

Figure Captions

Fig 1. ϕ-values of acylphosphatase (AcP). Experimentally measured ϕ-values[4] and calculated ϕ-values obtained from conformation samples properly weighted with respect to the uniform distribution in Ω_{ϕ}.

Fig 2. Defining folded and unfolded states and selecting conformations with 0.5 probability of folding. (a) The thresholds (vertical dashed lines) of the fraction of native contacts preserved for the folded and unfolded states. (b) Number counts of Markov chain Monte Carlo runs that reach the folded state for the set of 4,185 conformations. Each point represents one conformation. Only the conformations between the two horizontal lines are included in TSE.

Fig 3. The distributions of cRMSD values of conformations satisfying the ϕ-value constraints. The distributions for (a) the transitions state ensemble Ω_{TSE} of 1,501 clusters of conformations, (c) the denatured-side ensemble Ω_{DS}, and (e) the native-side ensemble Ω_{NS} to the native conformation of protein acylphosphatase at different cRMSD distance intervals, and the distributions of the fraction of native contacts preserved at different intervals for (b) Ω_{TSE}, (d) Ω_{DS}, and (f) Ω_{NS}. Both unweighted (white bar) and weighted (solid gray) distributions are shown.

Fig 4. Lack of correlation between energy and cRMSD of conformations in the TSE.

Fig 5. The distributions of cRMSD between the secondary structures in the weighted TS conformations and in the native state of protein acylphosphatase. (a) Helix α_1 (residues 22-33, white bar) and helix α_2 (residues 55-66, gray bars); (b) Strand β_3 (residues 46-53,

white bar), strand β_4 (residues 77-85, gray bars).

Fig 6. The recovery of overall ϕ-values and resulting larger cRMSD of conformations generated with ϕ-values constrained only at three key residues of Y11, P54, and F94. (a) Experimentally measured ϕ-values and calculated ϕ-values obtained from conformation samples satisfying the ϕ-value constraints of three key residues. (b) The distributions of cRMSD values of conformations satisfying the ϕ-value constraints of three key residues (white bar) and 24 residues (solid gray).

Fig 7. The average point-wise distances of residues between the weighted TSE and the native conformation of protein acylphosphatase. The three circles are the three key residues identified by Vendruscolo et al[40] that have large experimentally measured ϕ-values. They have small point-wise cRMSD values.

Fig 8. The fractions of preserved native contacts with different sequence separations of protein acylphosphatase for (a) the weighted TSE, and (b) the native conformation. Bin 1-11 correspond to sequence separations of 4, 5, 6–10, 11–20, 21–30, 31–40, 41–50, 51–60, 61–70, 71–80, and 81–90, respectively.

Fig 9. Estimating the first passage time (FPT) to folded structure and correlation between FPT and fraction of native contacts among TSE. (a) An illustration of counting the first passage time $\widetilde{\xi}_{ji}$ as $t_1 - t(j)$. (b) The average FPT of conformations in TSE of AcP. Each point represents a transition state conformation.

Fig 10. The agreement of counted and calculated traveling times between states. For the fixed number of residues $L = 11$ and the end to end distance $r = 5\,\text{Å}$, this figure shows: (a) The frequency of different states in the trajectory of the MD simulation; and the ratio of the counted traveling time to the calculated traveling time for fixed destination state (b) 5, (c) 40, and (d) 75. Except for rarely observed states, counted and calculated traveling times

agree well with each other.

Fig 11. First passage time to folded structures and distance in cRMSD to the native structure. (a) The average first passage time of transition state conformations with different cRMSD distance to the native structure. For comparison, (b), (c), and (d) plot the distributions of the first passage time of the conformations in Ω_{TSE}, Ω_{DS}, and Ω_{NS}, respectively. Both unweighted (white bar) and weighted (solid gray) distributions are shown.

Fig 12. Effects of different ordering of conformations on clustering. (a) The distributions of cRMSD values of representative conformations of clusters obtained when conformations are ordered by weights (white bar) and when they are randomly ordered (solid gray). (b) The distributions of fractions of native contacts for conformation clustering obtained using conformations ordered by weights (white bar) and using random ordered conformations (solid gray). Overall, these distributions are similar.

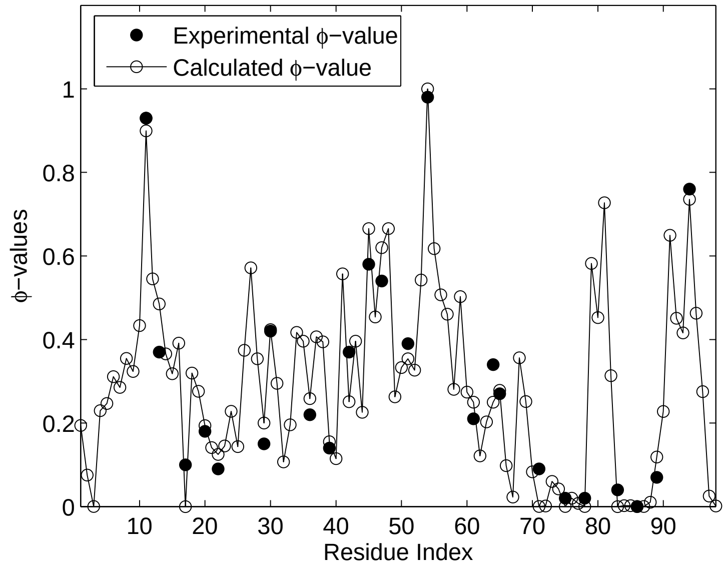
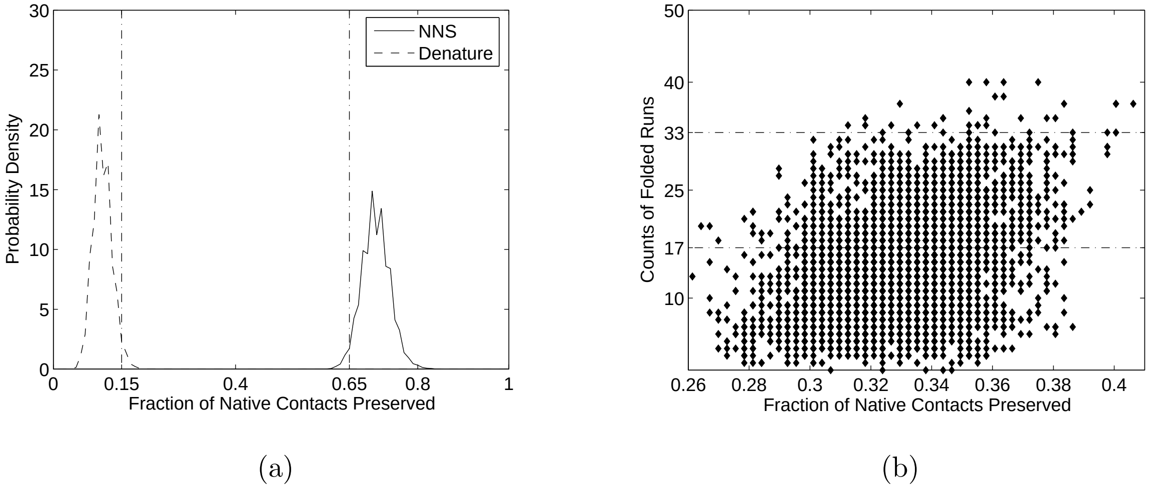

Experimental ϕ−value
Calculated ϕ−value
ϕ−values
1
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60
70
80
90
Residue Index

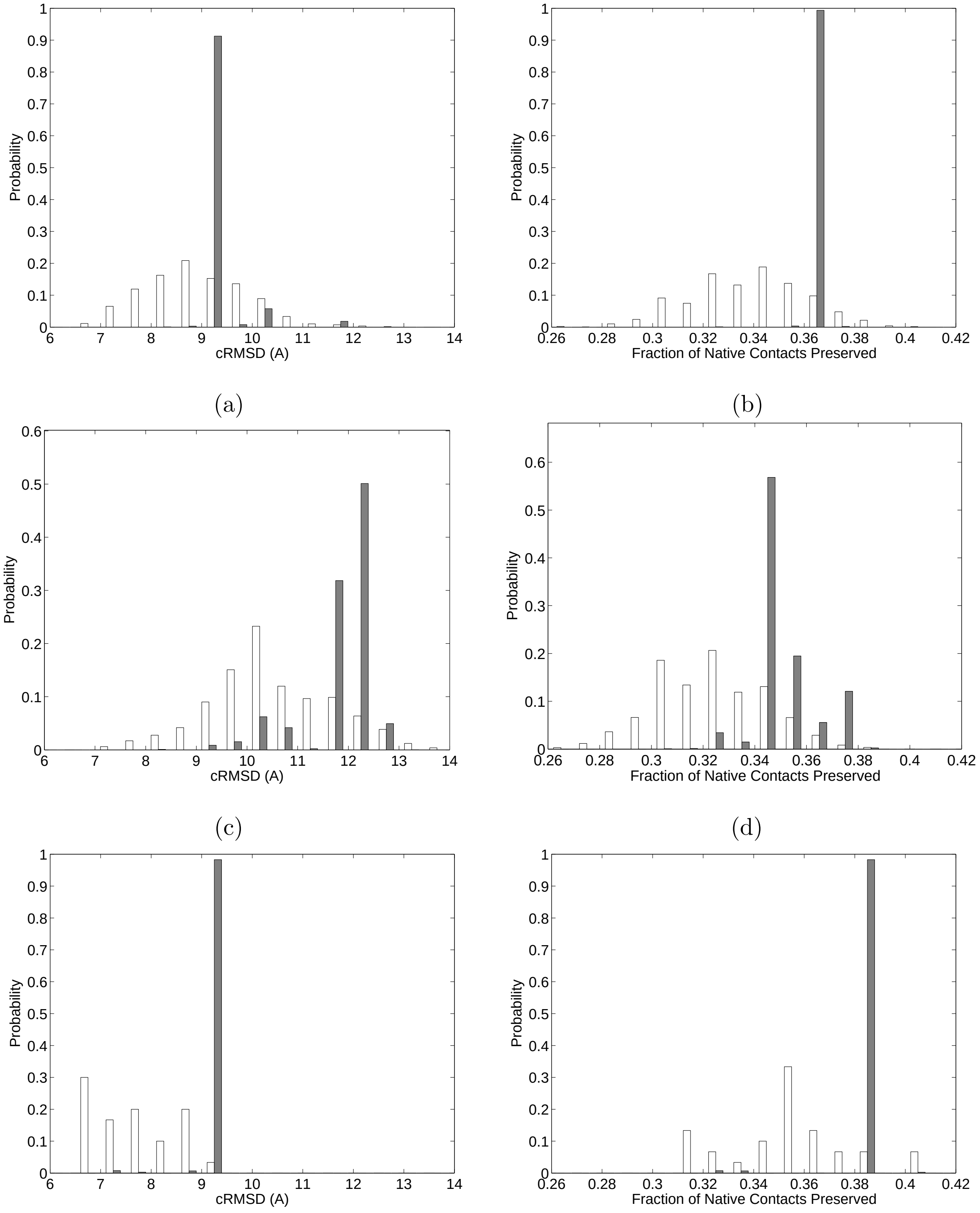
FIG. 1:


30
25
20
15
10
5
0
Probability Density
NNS
Denature
0
0.15
0.4
0.65
0.8
1
Fraction of Native Contacts Preserved
50
40
33
25
17
10
Counts of Folded Runs
0.26
0.28
0.3
0.32
0.34
0.36
0.38
0.4
Fraction of Native Contacts Preserved

(a)

(b)

FIG. 2:


Probability
cRMSD (A)
6
7
8
9
10
11
12
13
14
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Fraction of Native Contacts Preserved
0.26
0.28
0.3
0.32
0.34
0.36
0.38
0.4
0.42

(a)

(b)

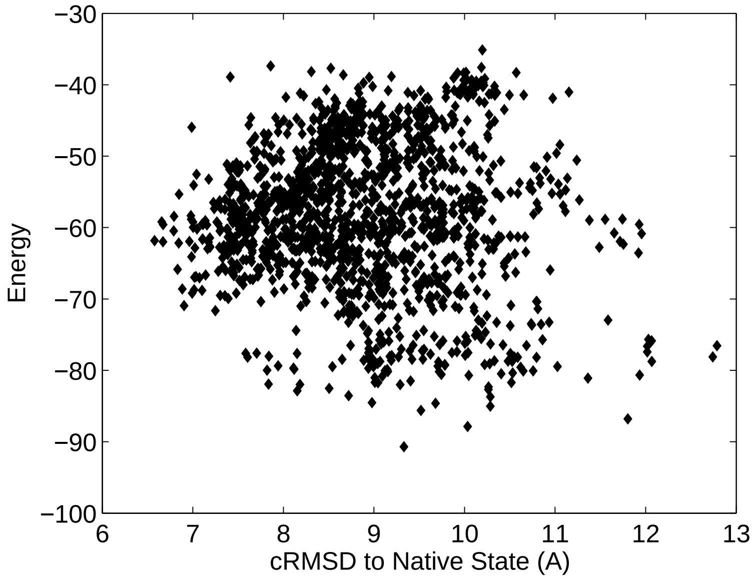
(c)

(d)

(e)

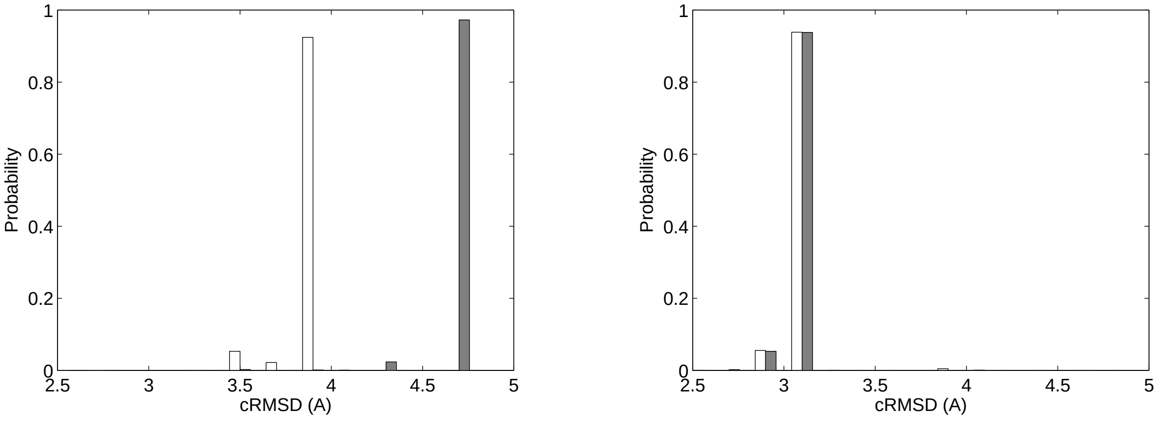
(f)

FIG. 3:



−30
−40
−50
−60
−70
−80
−90
−100
Energy
6
7
8
9
10
11
12
13
cRMSD to Native State (A)

FIG. 4:

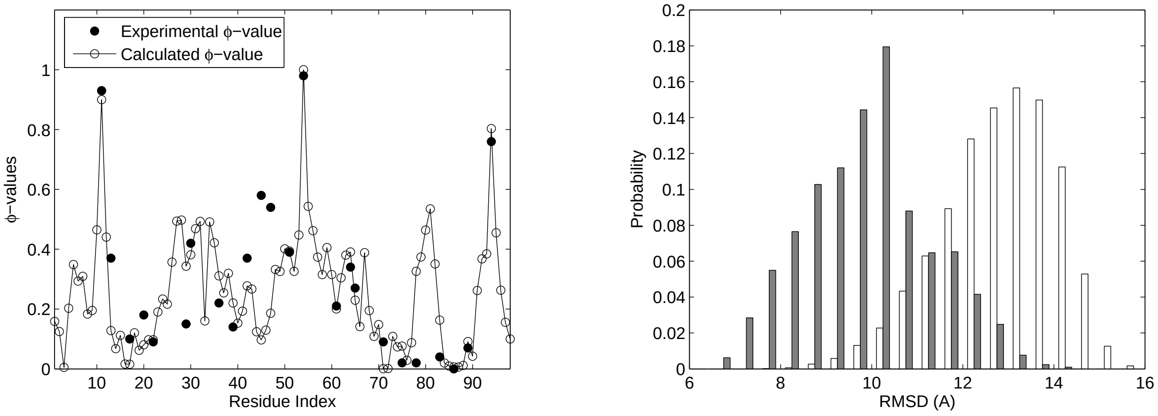


1
0.8
0.6
0.4
0.2
0
Probability
2.5
3
3.5
4
4.5
5
cRMSD (A)
1
0.8
0.6
0.4
0.2
0
Probability
2.5
3
3.5
4
4.5
5
cRMSD (A)

(a)

(b)

FIG. 5:


Experimental ϕ−value
Calculated ϕ−value
ϕ−values
1
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60
70
80
90
Residue Index
Probability
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
6
8
10
12
14
16
RMSD (A)

(a)

(b)

FIG. 6:

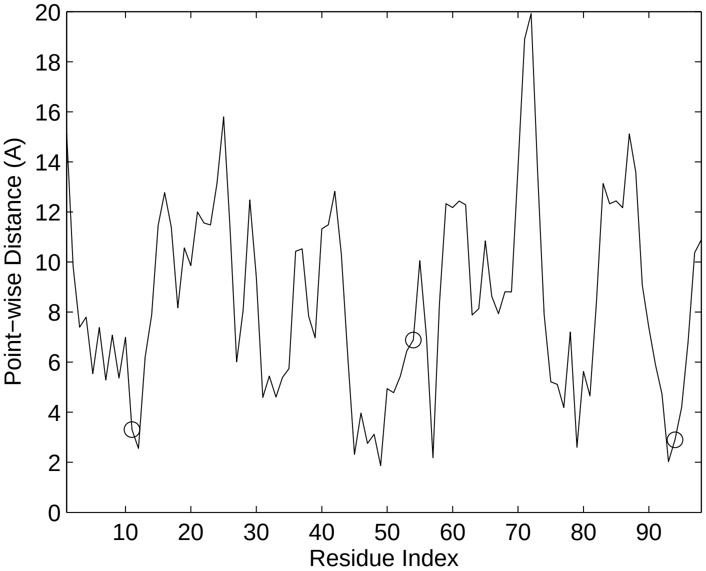
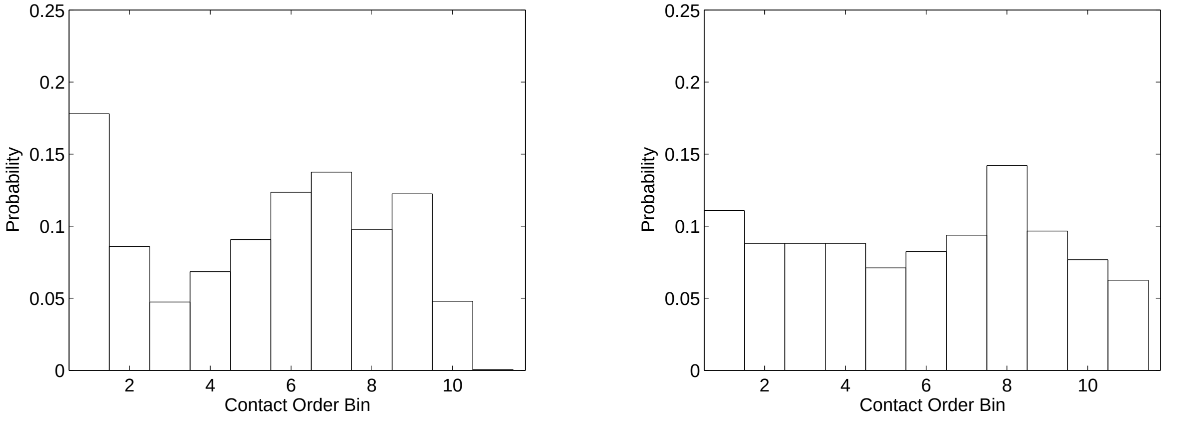

Point−wise Distance (A)
20
18
16
14
12
10
8
6
4
2
0
10
20
30
40
50
60
70
80
90
Residue Index

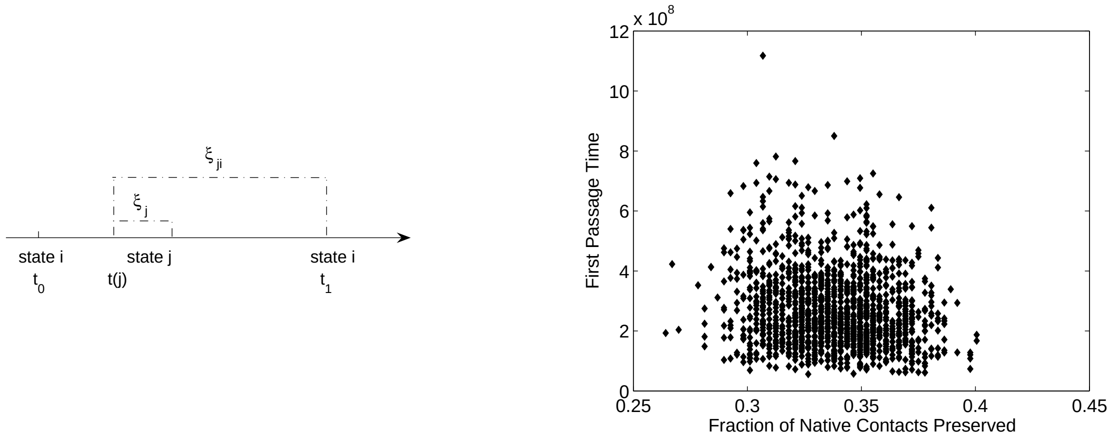
FIG. 7:



(a)

(b)

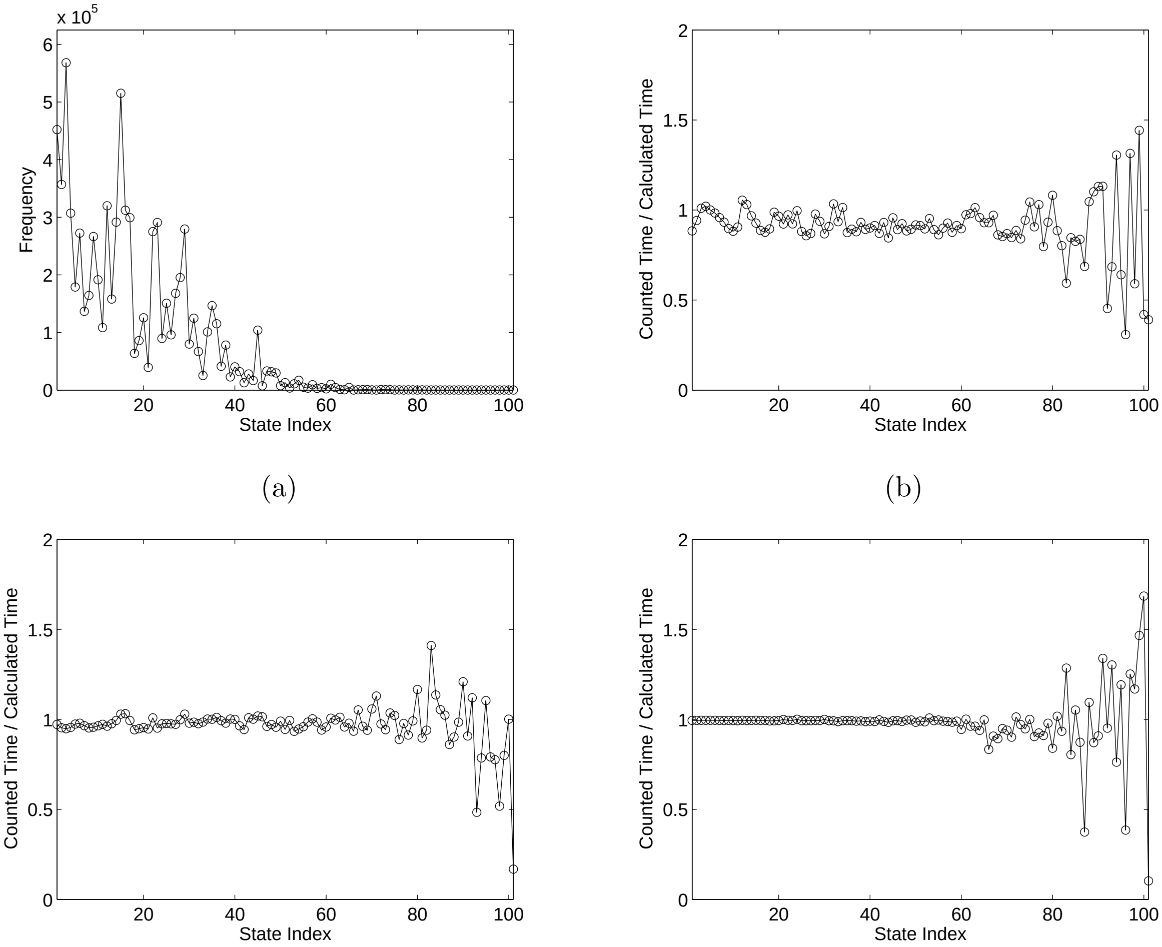
FIG. 8:



(a)

(b)

FIG. 9:


x 10^5
Frequency
State Index
Counted Time / Calculated Time
0
1
2
3
4
5
6
0.5
1.5
20
40
60
80
100

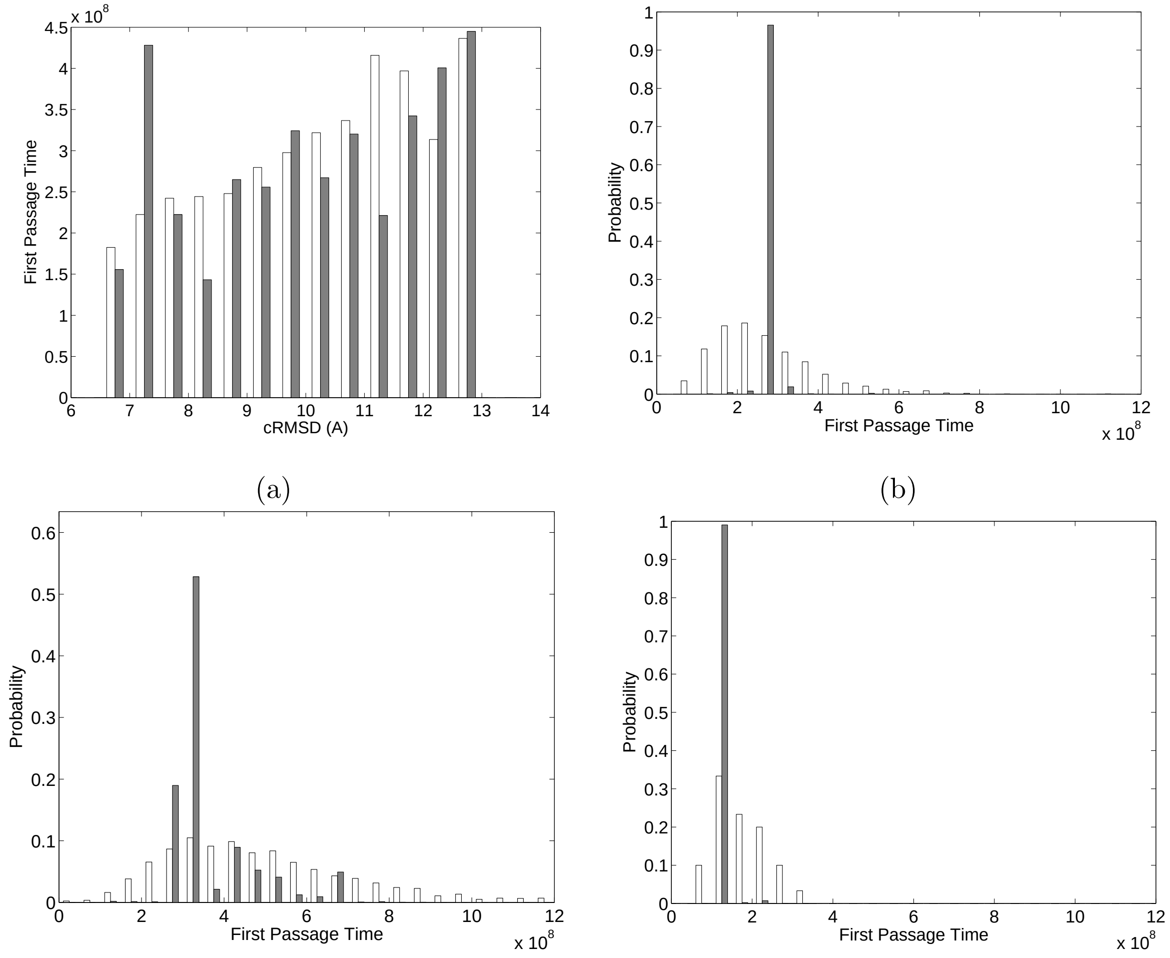
(a)

(b)

(c)

(d)

FIG. 10:


x 10^8
First Passage Time
cRMSD (A)
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5
6 7 8 9 10 11 12 13 14
Probability
First Passage Time
x 10^8
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
0 2 4 6 8 10 12
Probability
First Passage Time
x 10^8
0 0.1 0.2 0.3 0.4 0.5 0.6
0 2 4 6 8 10 12
Probability
First Passage Time
x 10^8
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
0 2 4 6 8 10 12

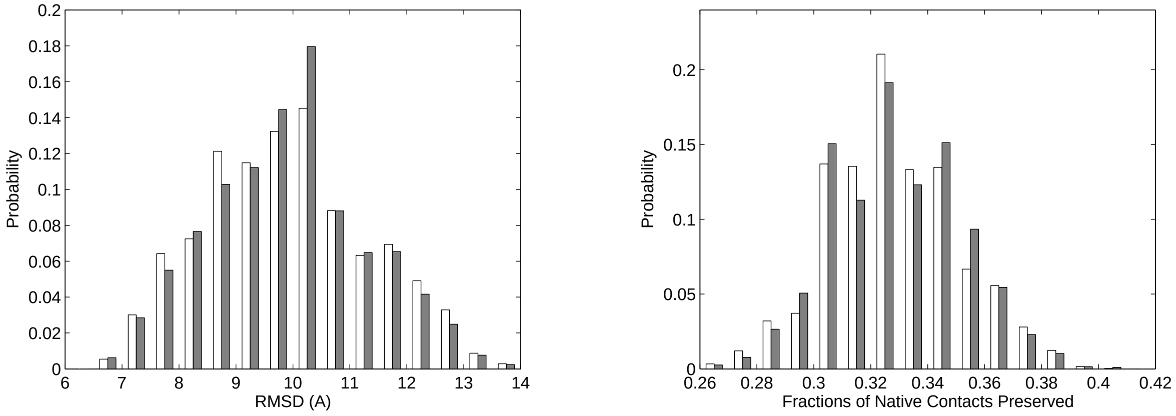
(a)

(b)

(c)

(d)

FIG. 11:


0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Probability
6
7
8
9
10
11
12
13
14
RMSD (A)
0.2
0.15
0.1
0.05
0
Probability
0.26
0.28
0.3
0.32
0.34
0.36
0.38
0.4
0.42
Fractions of Native Contacts Preserved

(a)

(b)

FIG. 12: