Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.
This journal is © the Owner Societies 2019

Supplementary data

Selective Enhanced Sampling in Dihedral Energy Term Facilitating to Overcome the Dihedral Energy Increase in Protein Folding and Accelerating the Searching for Protein Native Structure

Qiang Shao[1,2]*, Lijiang Yang[2,3]*, Weiliang Zhu[1,4]

[1]Drug Discovery and Design Center, CAS Key Laboratory of Receptor Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai, 201203, China

[2]Beijing National Laboratory for Molecular Sciences, 1st North Street, Zhongguancun, Beijing, 100080, China

[3]Institute of Theoretical and Computational Chemistry, College of Chemistry and Molecular Engineering, and Biodynamic Optical Imaging Center, Peking University, Beijing 100871, China

[4]Open Studio for Druggability Research of Marine Natural Products, Pilot National Laboratory for Marine Science and Technology, 1 Wenhai Road, Aoshanwei, Jimo, Qingdao, 266237, China

*To whom correspondence should be addressed. Qiang Shao, Tel: +86 21 50806600-1304, E-mail: qshao@simm.ac.cn; Lijiang Yang, E-mail: lijiangy@pku.edu.cn.

Computational Methods

Preliminary iteration process in P-SITSMD or D-SITSMD to determine the parameter values of n_k's. In P-SITSMD or D-SITSMD, to fulfill the thorough sampling in the desired energy range (potential or dihedral energy only), the energy must be sampled evenly in a large temperature range. The potential energy distribution function is:

$$p(E_{eff}) = e^{-\beta_0 E_{eff}} = \sum_k n_k e^{-\beta_k E} \quad (1)$$

The parameter values of individual terms of n_k's can be computationally determined in a preliminary iteration process to follow the requirement that each term in Eq. 1 contributes equally to the total distribution. We define $P_k = n_k \int_r e^{-\beta_k E_{eff}(r)} dr$ and preselect the ratios between P_k's for all *k* between 1 to *N*. Using the normalized quantities ($p_k = P_k / \sum_{k=1}^{N} P_k$), we then set fixed expectation values $\{p_k^0\}$. For the desired even distribution of P_k's, p_k is simply equal to *1/N*. In the preliminary iteration process, p_k can be calculated iteratively. Accordingly, n_k can be calculated and updated to the converged values according to the ratios between neighboring p_k values. Let's define a series of number m_k:

$$m_k = \begin{cases} 1 & k = 1 \\ n_{k+1}/n_k & 1 < k \leq N \end{cases} \quad (2)$$

The original series of n_k is related to m_k by $n_k = n_1 \prod_{i=1}^{k} m_i$. The preliminary iteration process in P-ITSMD/D-SITSMD is performed as follows:

1. A set of initial values of m_k's (m_k^0) are chosen to calculate the initial values n_k^0, which are then used to run preliminary MD simulation.

2. From the trajectory of preliminary MD, p_k are calculated ($p_k(0)$), a new set of m_k's are achieved

by $m_k(1) = m_k(0)\frac{p_k(0)}{p_{k+1}(0)}$, in order to fulfill the requirement of $\frac{p_k(1)}{p_{k+1}(1)} = 1$ in the next iteration step (even distribution over the energy range sampled). New set of values of n_k's are calculated from $m_k(1)$ and used in the next iteration step.

3. Repeating step 2 to update the values of m_k's as well as n_k's. From the second iteration step, history information of m_k's in earlier steps is involved in the calculation of new m_k's in i^{th} step:

$$m_k(i) = \frac{m_k(i-1)}{\sum_{l=0}^{i-1}\sqrt{p_k(l)p_{k+1}(l))}}\left[\sqrt{p_k(i-1)p_{k+1}(i-1)}\,\frac{p_k(i-1)}{p_{k+1}(i-1)} + \sum_{l=0}^{i-2}\sqrt{p_k(l)p_{k+1}(l))}\right] \qquad (3)$$

Until n_k's converge, the preliminary iteration process can be stopped. Taking the simulation system of chignolin as an example, a total of 80 n_k's were used in both P-SITSMD and D-SISTMD simulations. As shown in Figure S15, the values of n_k's are modified along the preliminary iteration process and finally go to the convergence for both simulations. Accordingly, the potential energy or dihedral energy can be sampled in an expanded range covering both low and high energy regions. Thus, 80 n_k's are enough for the enhanced sampling of the test simulation system, the overhead of using more n_k's is trivial. And the production run will not be affected after the converged n_k's are obtained.

Protein	Force Field	Simulation	N_{water}	System Size (Å^3)	Temp (K)	N_{nk}[a]
Chignolin	FF14SBonlysc	D-SITS	800	30.8×29.3×28.2	280-650	80
		P-SITS			280-650	80
		REMD			300-450	---
TC5b	FF14SBonlysc	D-SITS	2280	41.5×52.0×46.6	280-650	80
		P-SITS			280-650	80
TC5b	FF99SBildn	D-SITS	2350	41.5×52.0×48.4	280-650	80
		P-SITS			280-650	80
TC5b	FF03	D-SITS	2348	37.0×46.4×43.0	280-650	100
		P-SITS			280-650	80
HP35	FF14SBonlysc	D-SITS	3064	43.0×52.0×44.2	280-650	100

Table S1. Summary of the parameters for the protein systems under study. [a]N_{nk} represents the number of n_k's used in the SITS relevant simulations (the detailed values of individual terms of n_k's are not presented because of the large amount, which are automatically determined in a preliminary iteration process).

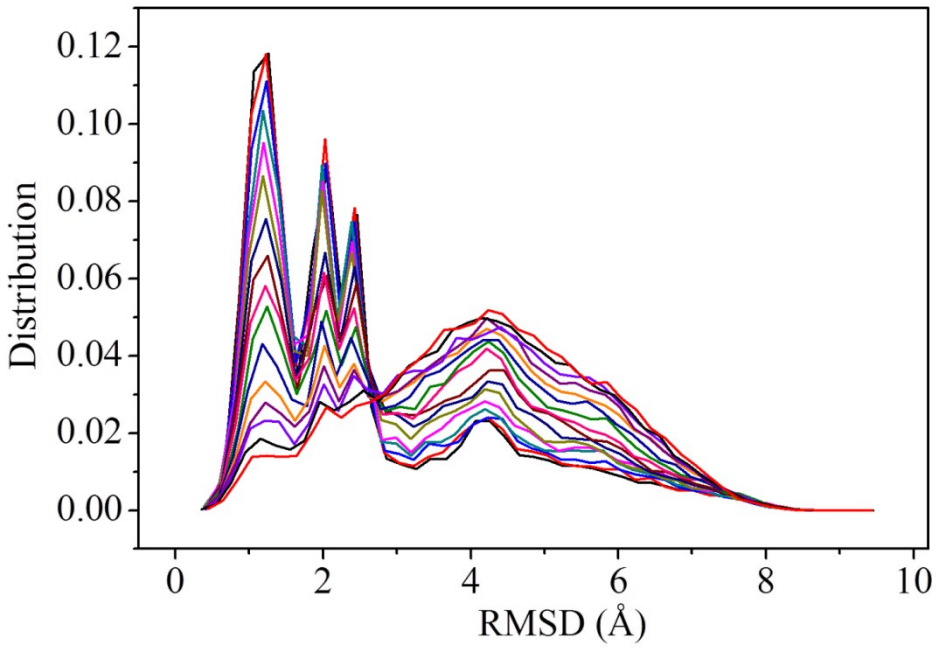

Figure S1. The distribution of root-mean-square deviation (RMSD) in individual replicas of REMD simulation of chignolin in water.

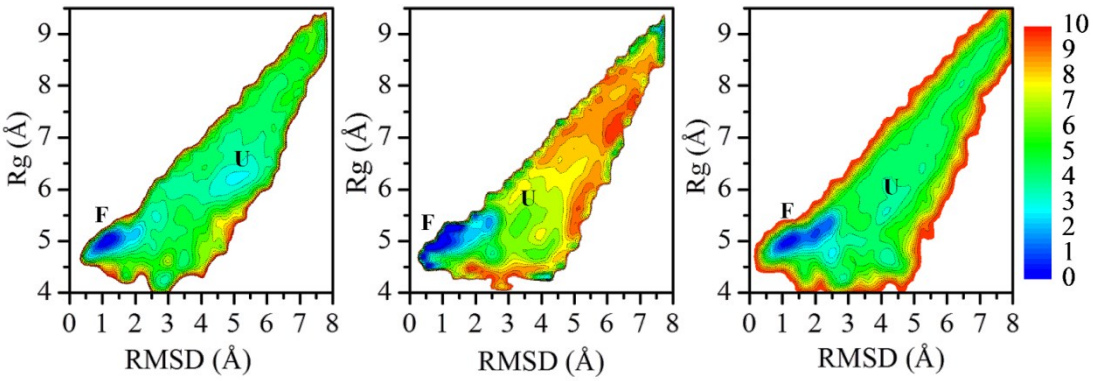

Figure S2. (Left to right) Two-dimensional free energy landscapes as the function of RMSD and R_g for the folding of chignolin simulated by D-SITSMD, P-SITSMD, and REMD simulations. The contours are spaced at intervals of 0.5 k_BT.

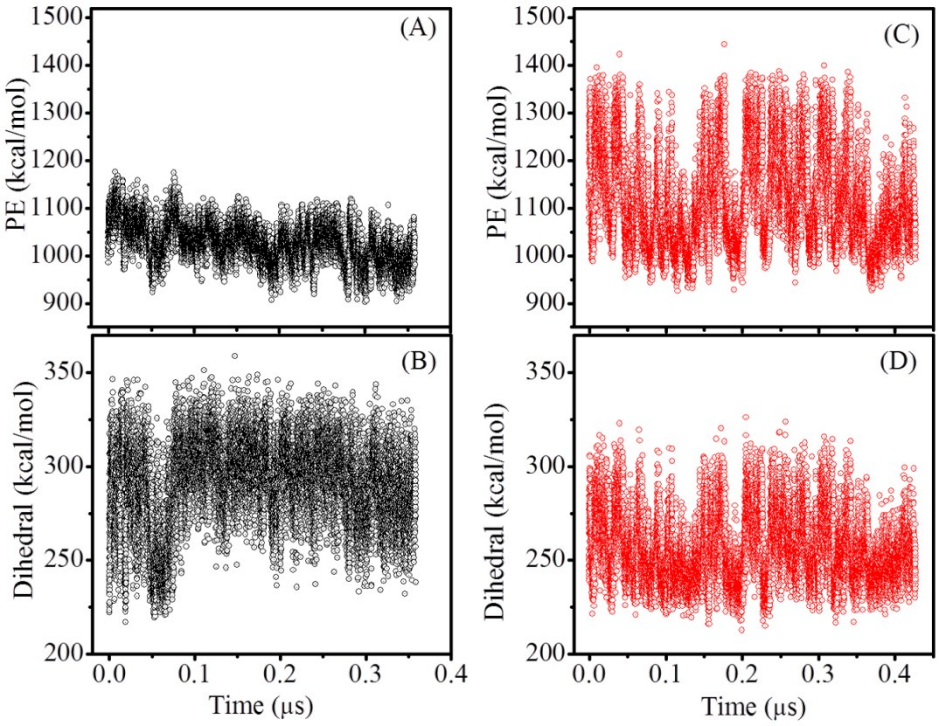

Figure S3. (A-B) Sampled ranges for potential energy (PE) and dihedral energy of protein in a representative trajectory of D-SITSMD simulation on TC5b modeled by FF14SBonlysc force field. (C-D) Sampled ranges for potential energy (PE) and dihedral energy of protein in a representative trajectory of P-SITSMD simulation on TC5b modeled by FF14SBonlysc force field.

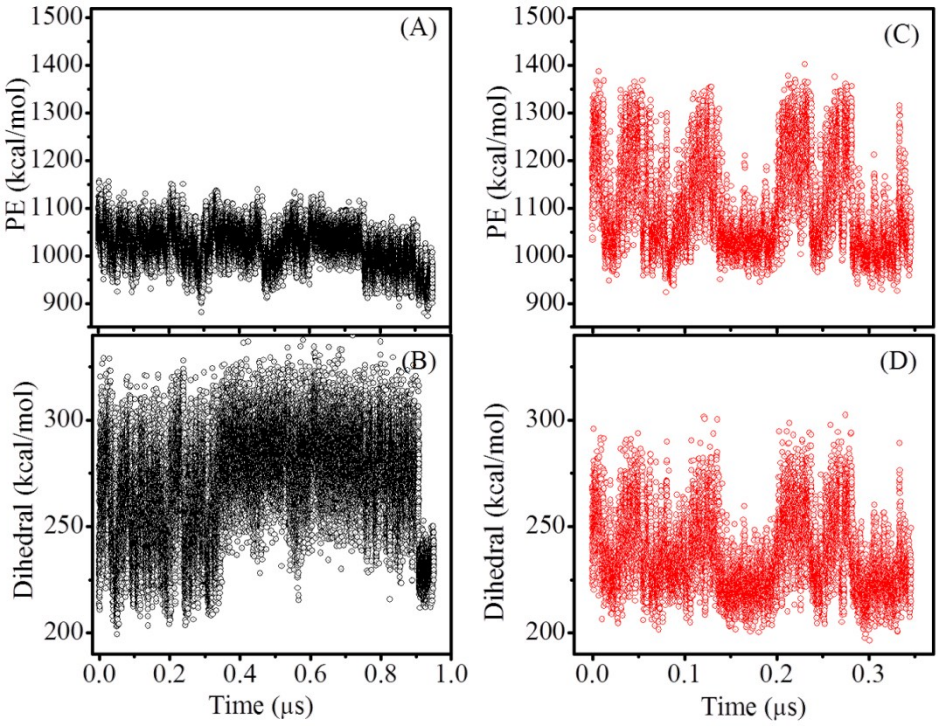

Figure S4. (A-B) Sampled ranges for potential energy (PE) and dihedral energy of protein in a representative trajectory of D-SITSMD simulation on TC5b modeled by FF99SBildn force field. (C-D) Sampled ranges for potential energy (PE) and dihedral energy of protein in a representative trajectory of P-SITSMD simulation on TC5b modeled by FF99SBildn force field.

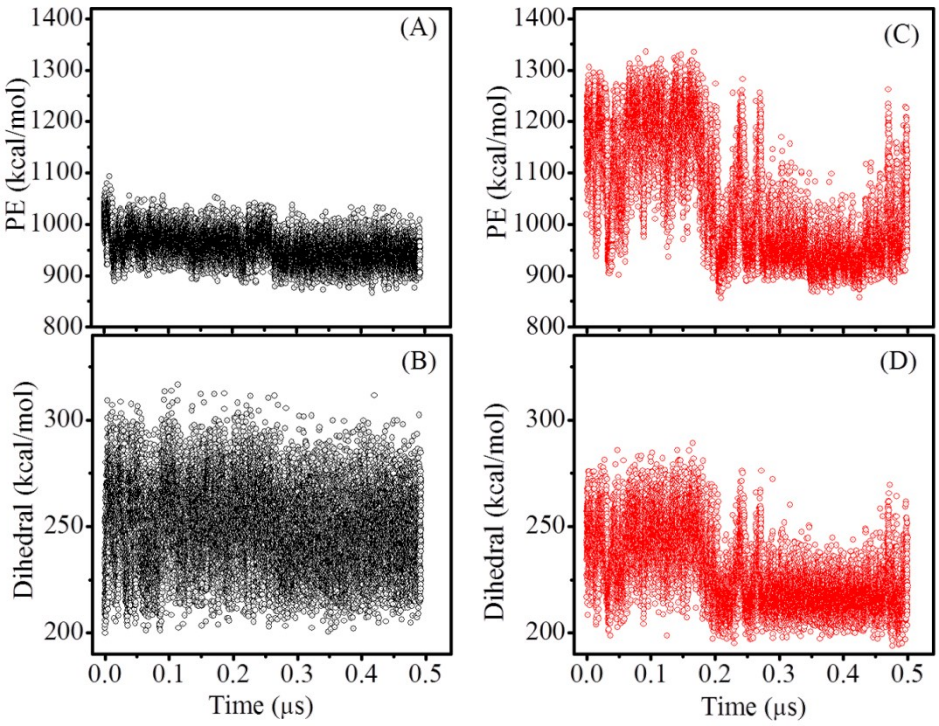

Figure S5. (A-B) Sampled ranges for potential energy (PE) and dihedral energy of protein in a representative trajectory of D-SITSMD simulation on TC5b modeled by FF03 force field. (C-D) Sampled ranges for potential energy (PE) and dihedral energy of protein in a representative trajectory of P-SITSMD simulation on TC5b modeled by FF03 force field.

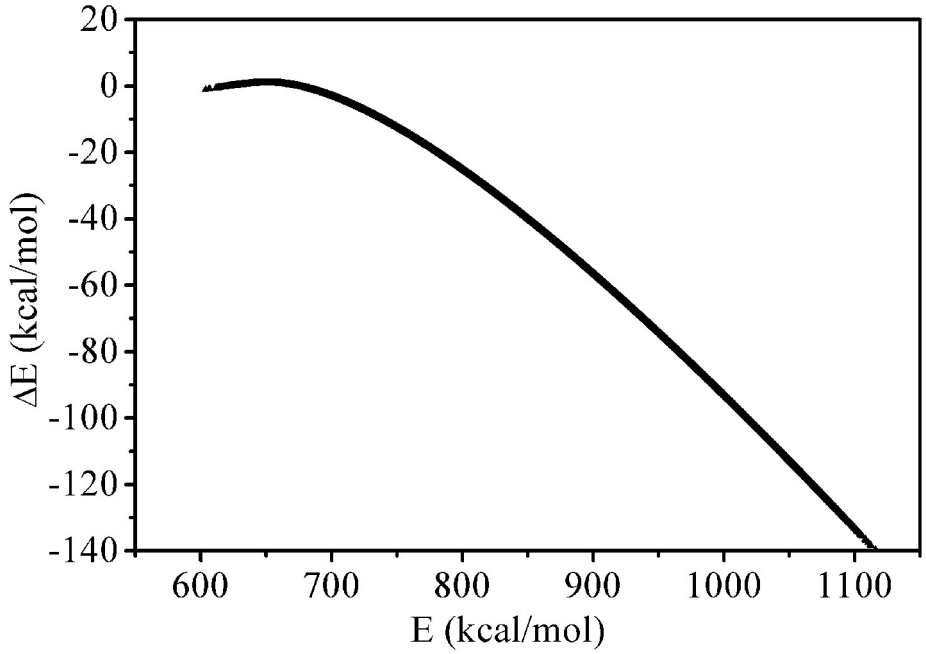

Figure S6. The correlation of the biased energies ($\Delta E = E_{eff} - E$) and the sampled potential energy (E) in the P-SITSMD simulation of FF14SBonlysc modeled TC5b system.

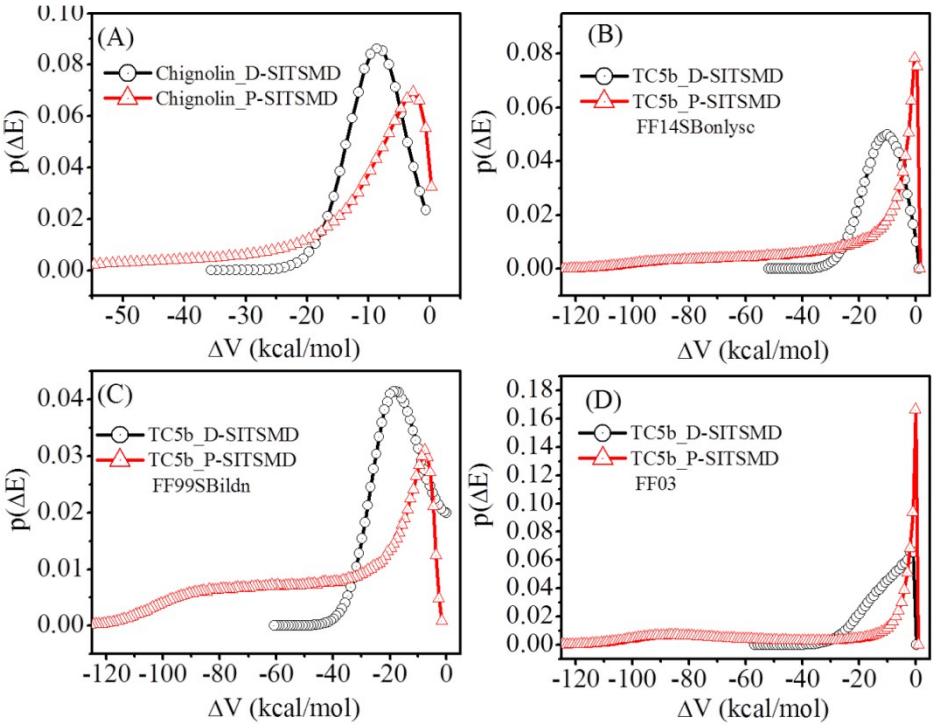

Figure S7. Distribution of the biased energies ($\Delta E = E_{eff} - E$) applied in the P-SITSMD and D-SITSMD simulations of (A) FF14SBonlysc modeled chignolin, (B) FF14SBonlysc modeled TC5b, (C) FF99SBildn modeled TC5b, and (D) FF03 modeled TC5b, respectively.

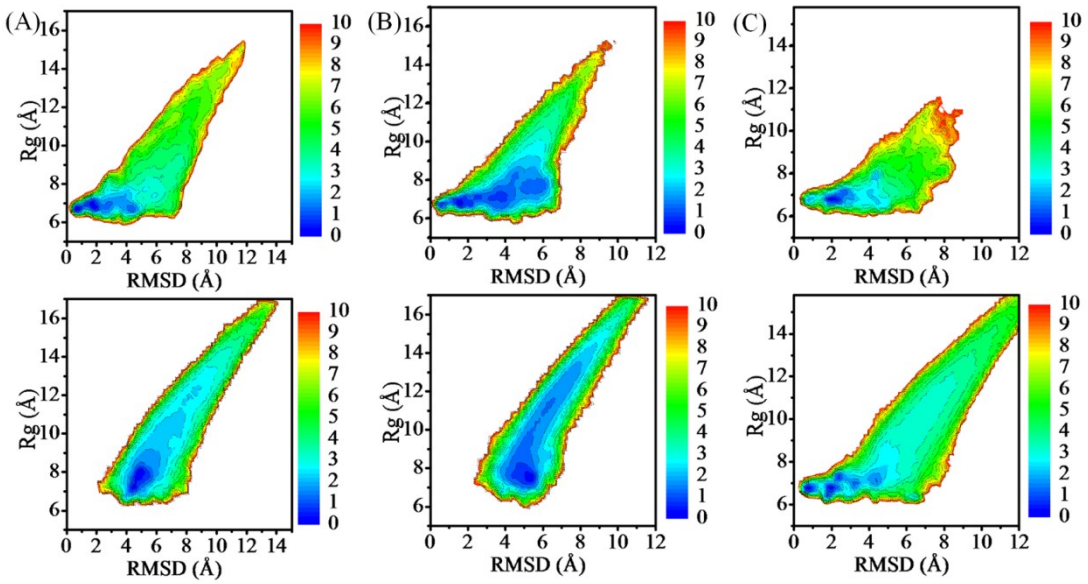

Figure S8. Two-dimensional free energy landscapes as the function of RMSD and R_g for the D-SITSMD (upper) and P-SITSMD (lower) simulations of TC5b with the force fields of (A) FF14SBonlysc, (B) FF99SBildn, and (C) FF03. The contours are spaced at intervals of 0.5 k_BT.

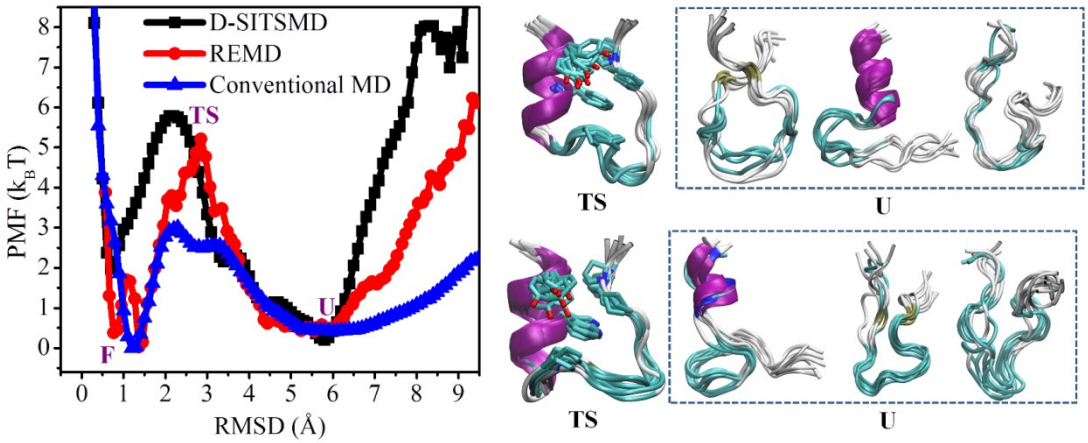

Figure S9. Left: Comparison of the potential of mean force (PMF) profiles along the coordinate of RMSD among the present D-SITSMD simulation and previous REMD simulation by Yu *et al.* for FF99SBildn modeled TC5b, and the extensive conventional MD simulation by Lindorff-Larsen *et al.* for CHARMM22* modeled TC10b_K8A. Right: The representative (populated) conformations of the unfolded (U) and transition state ensemble (TS) states are presented (the present results at the top and the results by Lindorff-Larsen *et al.* at the bottom).

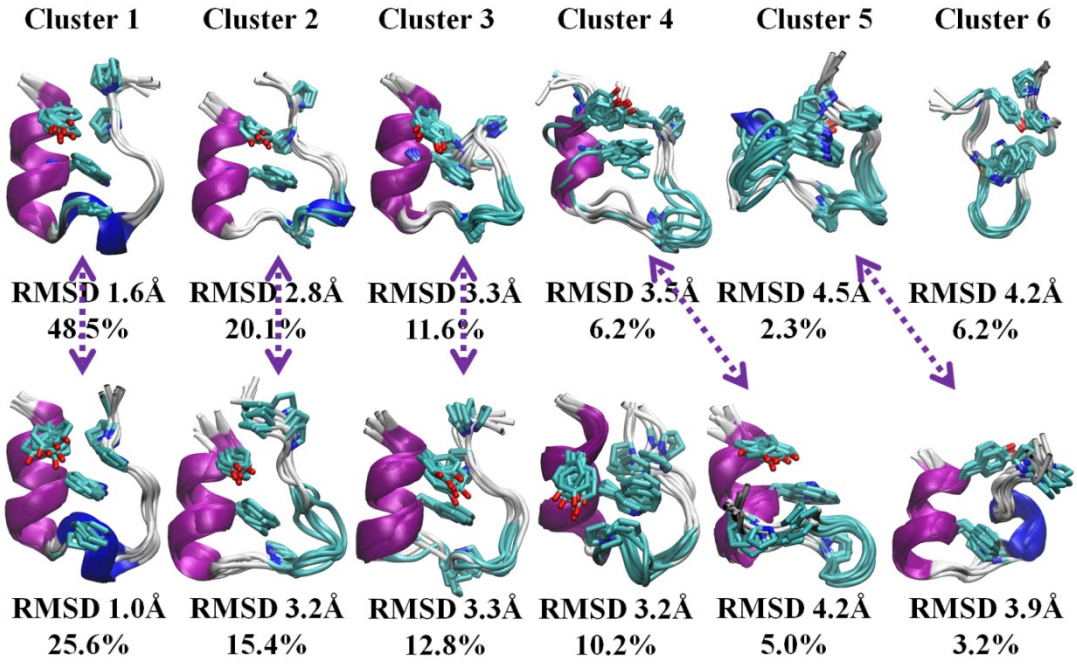

Figure S10. The six most populated clusters in the equilibrium conformational ensembles of TC5b. The clusters measured from the present D-SITSMD simulation are shown at the top and the clusters from the standard dual-boost AMD at the bottom. Under the representative structure of each cluster is shown the RMSD and the average population in the cluster ensemble. The roughly similar clusters observed in both simulations are connected by purple dashed lines.

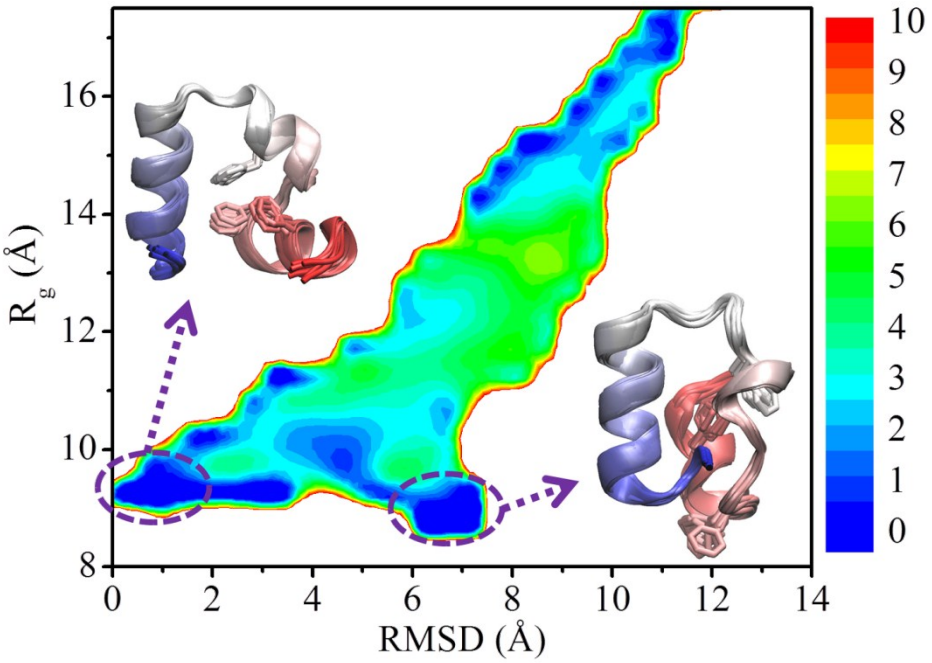

Figure S11. Two-dimensional free energy landscape as the function of RMSD and R_g for the D-SITSMD simulation of HP35, along with the representative conformations of the basins corresponding to the folded and unfolded states. The contours are spaced at intervals of 0.5 k_BT.

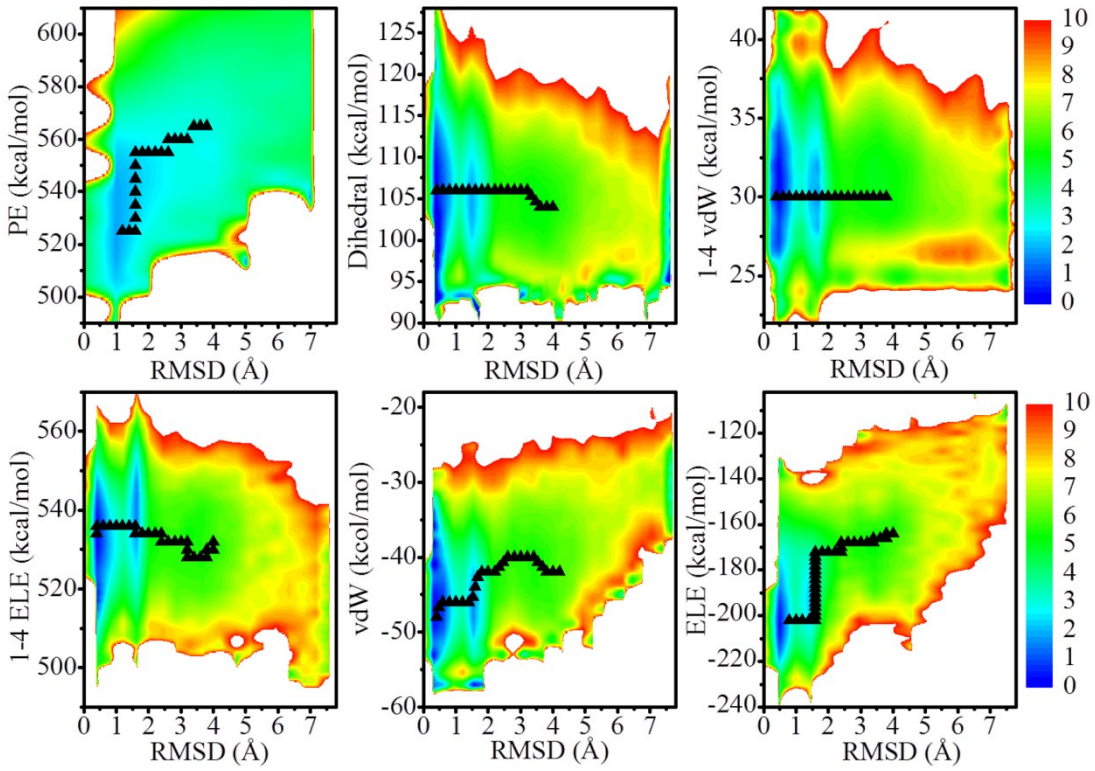

Figure S12. Two-dimensional free energy landscapes as the function of RMSD and various energy components for the folding of chignolin simulated by P-SITSMD. The contours are spaced at intervals of 0.5 k_BT.

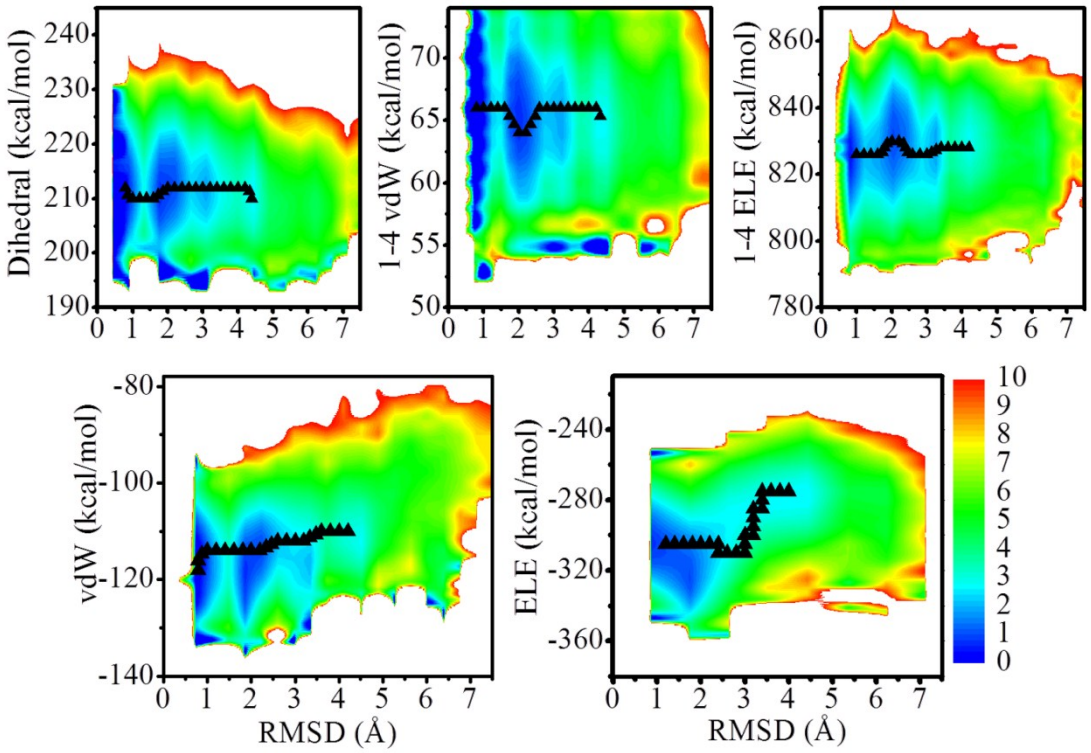

Figure S13. Two-dimensional free energy landscapes as the function of RMSD and various energy components for the folding of TC5b simulated by P-SITSMD and FF03 force field. The contours are spaced at intervals of 0.5 k_BT.

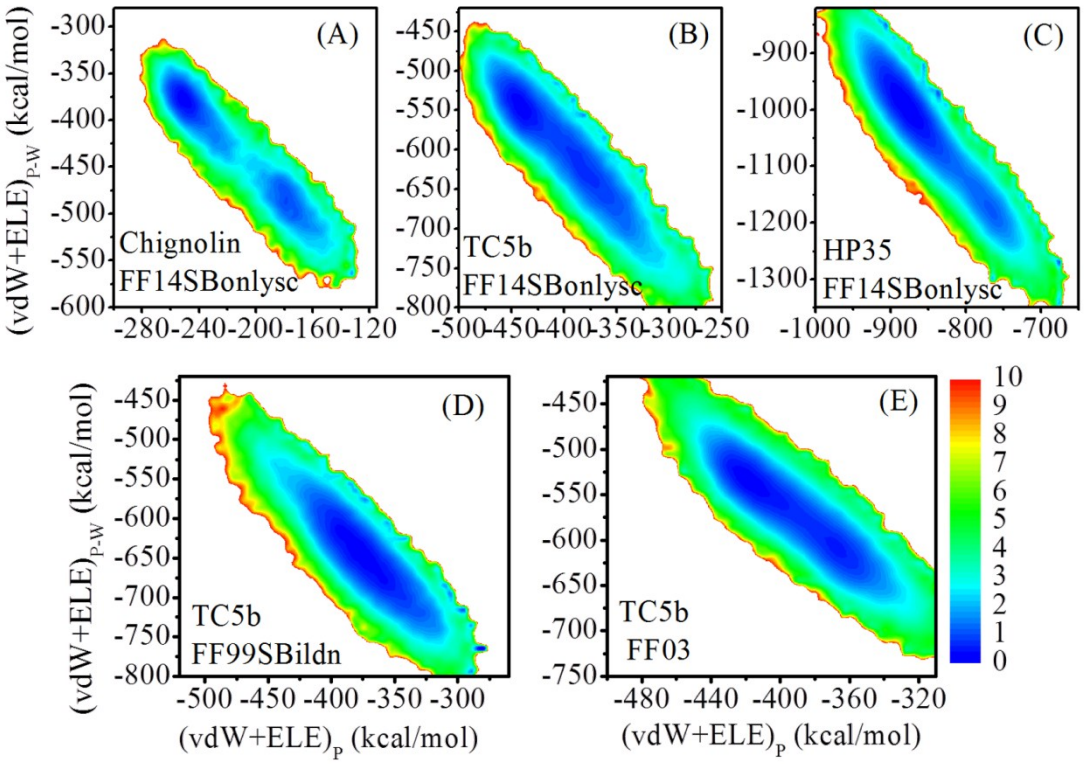

Figure S14. Two-dimensional free energy landscapes as the function of the combined vdW and electrostatic energies inside protein and between protein-water for (A) chignolin modeled by FF14SBonlysc, (B) TC5b modeled by FF14SBonlysc, (C) HP35 modeled by FF14SBonlysc, (D) TC5b modeled by FF99SBildn, and (E) TC5b modeled by FF03, respectively. The contours are spaced at intervals of 0.5 k_BT.

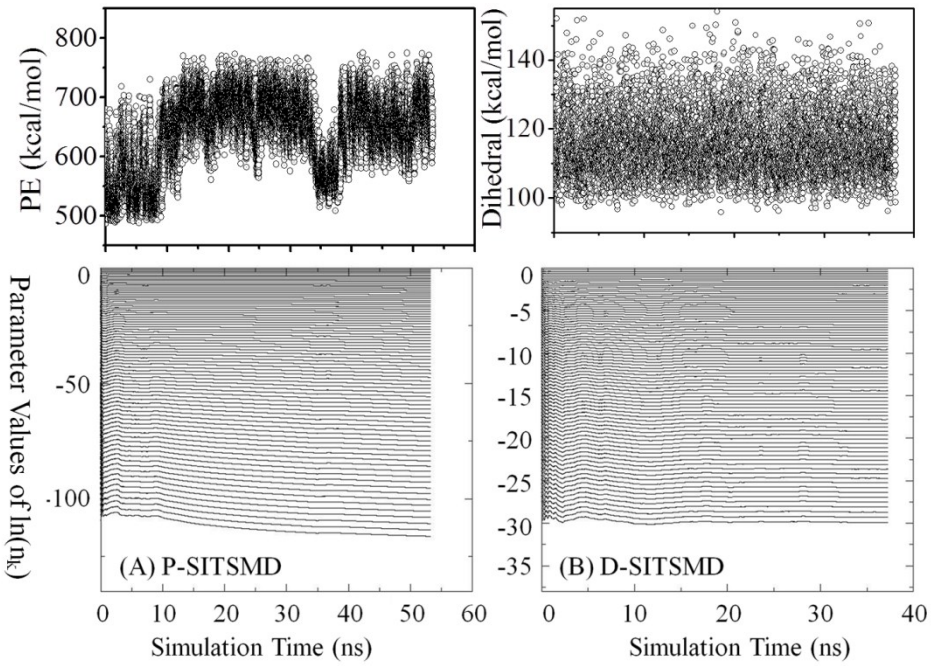

Figure S15. The sampled energies (top) and the variance of the values of individual terms of n_k's (bottom) along the preliminary iteration processes for (A) P-SITSMD and (B) D-SISTMD simulations of chignolin in water.