Supporting Information: Simulating Large-scale Conformational Changes of Proteins by Accelerating Collective Motions Obtained from Principal Component Analysis

Junhui Peng, Zhiyong Zhang*

Hefei National Laboratory for Physical Science at Microscale and School of Life Sciences, University of Science and Technology of China, Hefei, Anhui 230026, People's Republic of China

*Corresponding author: Zhiyong Zhang, Tel: +86-551-63600854; Fax: +86-551-3600374; Email: zzyzhang@ustc.edu.cn

MM/GBSA method

The MM/GBSA (molecular mechanics generalized Born surface area) is a semi-empirical method to quickly calculate the absolute free energy of a protein in solution, which combines the molecular mechanical energies and the continuum solvent models.[1] In the framework of MM/GBSA, the free energy is computed over an ensemble of structures:

$$\langle G \rangle = \langle E_{MM} \rangle + \langle G_{solv} \rangle - T \langle S_{solute} \rangle, \quad \text{(S1)}$$

where E_{MM} is the molecular mechanical term that includes bond, angle, torsion angle, van der Waals, and electrostatic energies. G_{solv} is the salvation free energy that consists of the polar and non-polar contributions:

$$G_{solv} = G_p + G_{np}. \quad \text{(S2)}$$

The polar term is calculated using the generalized Born model,[2] and the non-polar contribution is proportional to the solvent accessible surface area (SASA):

$$G_{np} = \gamma SASA + b. \quad \text{(S3)}$$

S_{solute} is the configurational entropy of the protein, which is usually much smaller than other terms. Therefore the solute entropy is not considered in our calculations.

We used the "-rerun" option of the program "mdrun" in the GROMACS-4.5.5 package,[3] to compute the MM/GBSA free energy. The xtc-format trajectory file, which contains only the coordinates of the protein, was chosen as the structure ensemble. The Charmm27 force field[4] was used to calculate the molecular mechanical energy E_{MM}. In the mdp file, the option of "GBSA" was turned on. The

Onufriev-Bashford-Case (OBC) method[2] was chosen to calculate the Born radii, and the scale factors were set as their default values. The surface tension was 2.25936 kJ mol^{-1} nm^{-2} for the non-polar part of the solvation free energy. After the "-rerun" calculation, the potential energies extracted from the energy trajectory file were actually the free energies without considering the solute entropy, and we then took the average.

Root mean square inner product between eigenvector sets

The similarity between any two sets of eigenvectors with the same dimensionality (obtained from PCA or ENM) can be measured by computing their root mean square inner product (RMSIP):

$$RMSIP = \left(\frac{1}{n_{ED}} \sum_{p=1}^{n_{ED}} \sum_{q=1}^{n_{ED}} \left(\Phi_p \cdot \Psi_q \right)^2 \right)^{1/2}, \tag{S4}$$

where Φ_p and Ψ_q are the two sets of eigenvectors, and n_{ED} is the number of the essential modes to be compared. The RMSIP value equals to 1 if the two eigenvector sets are identical.

Root mean square fluctuations from the ENM modes

From the low-frequency ENM modes, we can compute the root mean square fluctuation (RMSF) of each CG site in the protein by

$$\left(\left\langle \left(\Delta \mathbf{R}_I^{ED} \right)^2 \right\rangle \right)^{1/2} = \left(k_B T \sum_{h=1}^{3} \sum_{q=1}^{n_{ED}} \Psi_q^{I_h} \lambda_q^{-1} \Psi_q^{I_h} \right)^{1/2}, \tag{S5}$$

where k_B is the Boltzmann constant, and T is the absolute temperature. $\Psi_q^{I_h}$ is the h

component of CG site I in the ENM mode Ψ_q, and λ_q is the corresponding eigenvalue that is related to the frequency of the mode. n_{ED} is the number of the ENM modes used to calculate the RMSF.

References

(1) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712-3720.

(2) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383-394.

(3) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory. Comput.* **2008**, *4*, 435-447.

(4) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586-3616.

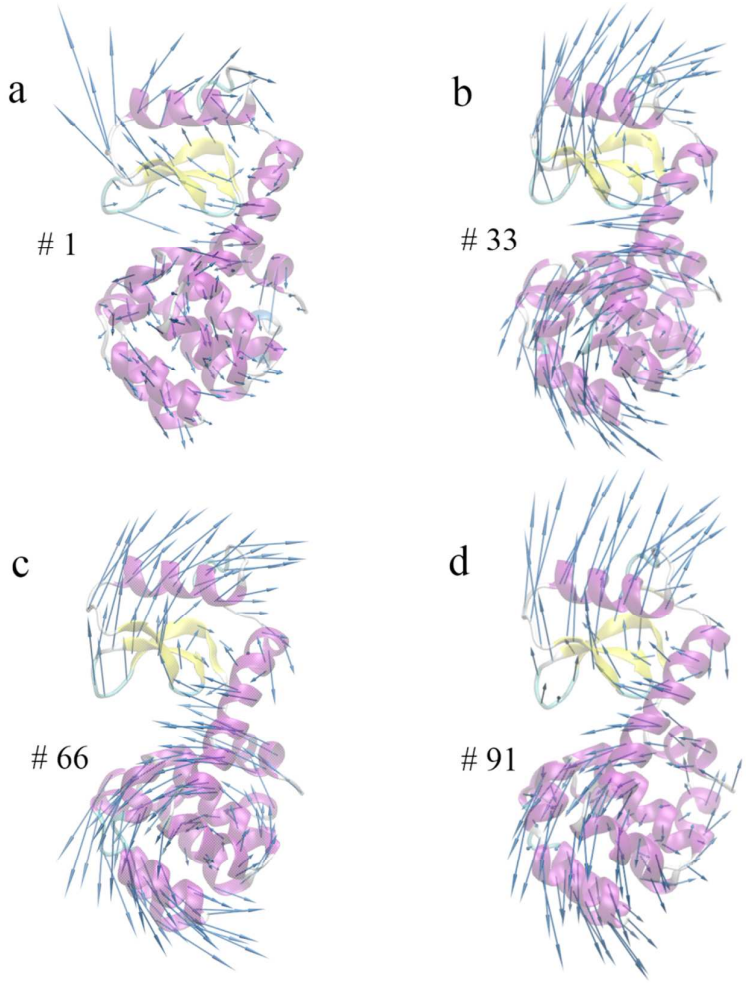

Figure S1. The largest-amplitude PCA mode of T4L obtained from several short trajectories, respectively, during the ACM-PCA simulation. "# 1" means the first simulation segment.

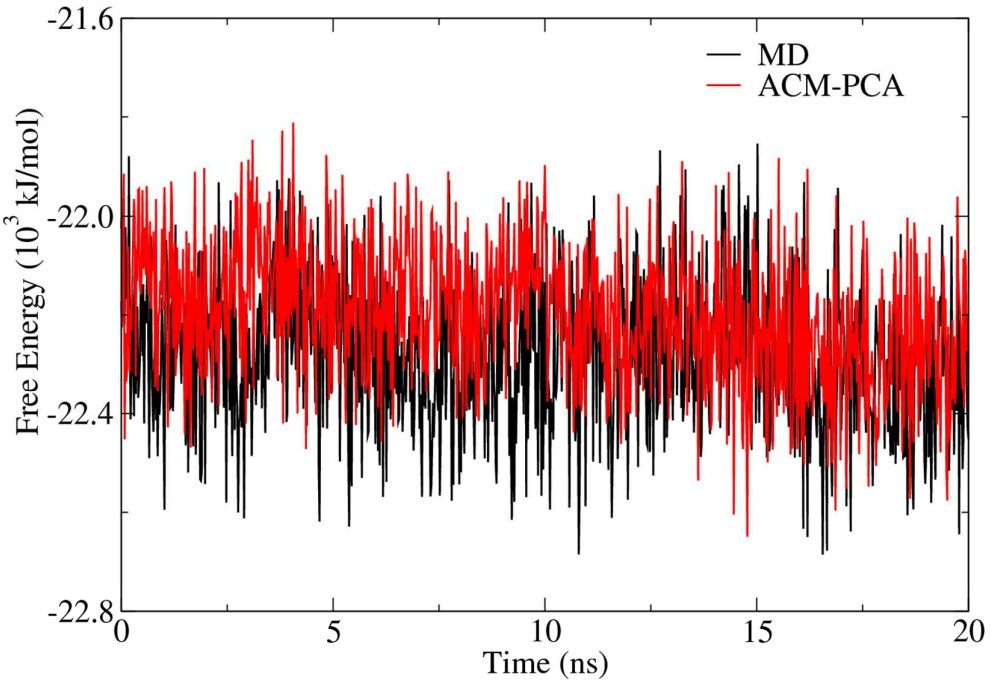

Figure S2. Absolute free energies of the conformations in the MD (black solid line) and ACM-PCA (red solid line) simulation, respectively, estimated by the MM/GBSA method.

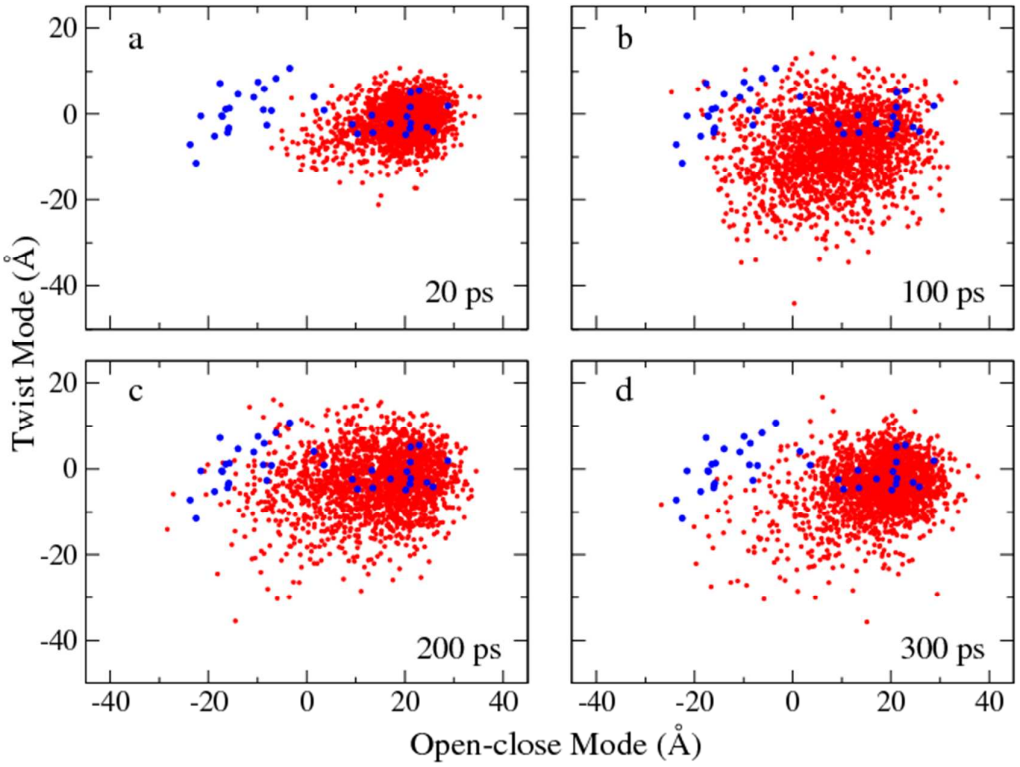

Figure S3. The ACM-PCA simulations of T4L with different lengths of simulation segment, such as (a) 20 ps, (b) 100 ps, (c) 200 ps, and (d) 300 ps. All the simulated conformations are projected onto the 2D essential subspace (colored by red), in order to show the sampling efficiency. The 38 experimental structures are also projected (colored by blue).

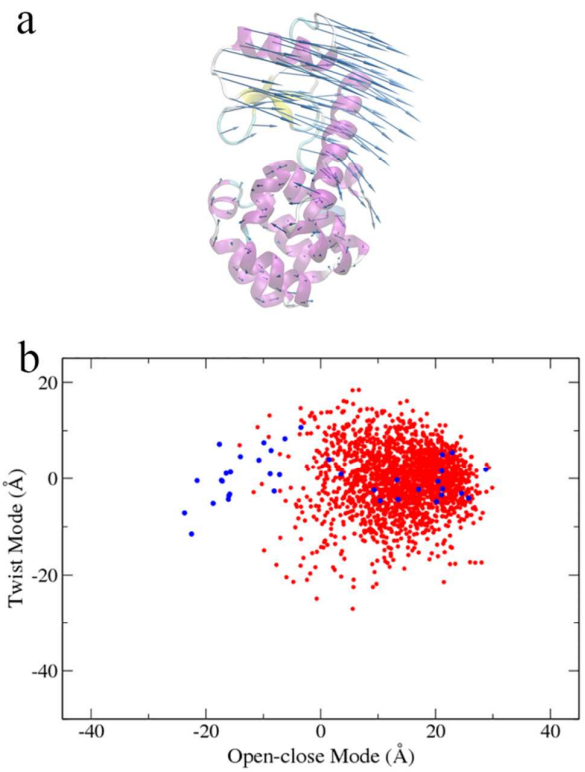

Figure S4. The ACM-PCA simulation of T4L using the following scheme of structure superposition, that is, all the conformations were superimposed by the C-terminal domain (COMs of the residues 75-162) only, but PCA was performed on the whole protein. (a) The first PCA mode obtained from one of the short simulations, and (b) all the simulated conformations are projected onto the 2D essential subspace (colored by red), in order to show the sampling efficiency. The 38 experimental structures are also projected (colored by blue).

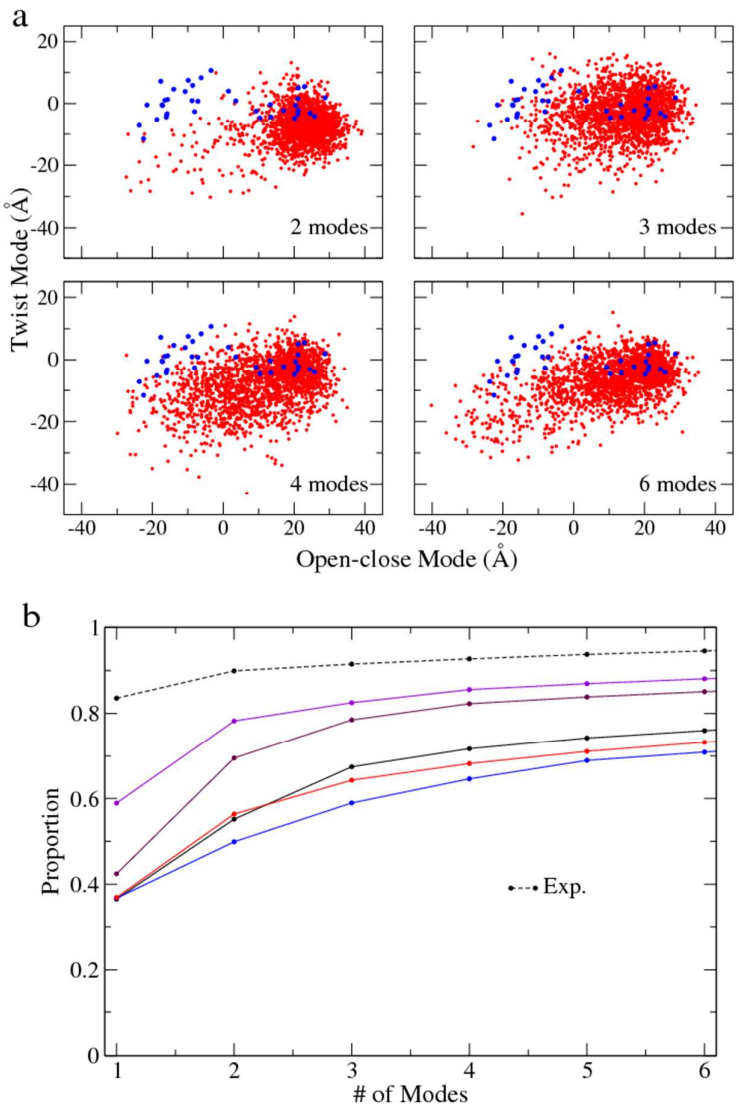

Figure S5. (a) The ACM-PCA simulations of T4L with different number of PCA modes (two, three, four, and six, respectively) coupled to 800 K. (b) Contribution of the PCA modes to the total fluctuation of T4L. The curve from PCA on the 38 experimental structures is indicated, and other curves are from PCA on several short trajectories.

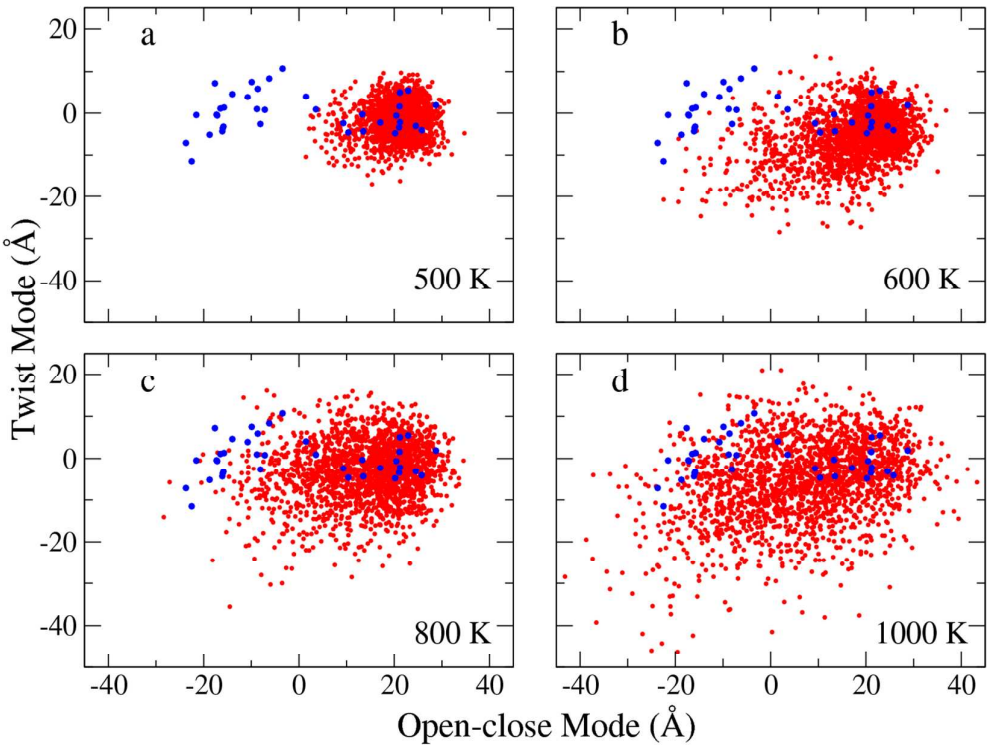

Figure S6. The ACM-PCA simulations of T4L, using different high temperatures to accelerate the collective domain motions. (a) 500 K, (b) 600 K, (c) 800 K, and (d) 1000 K.

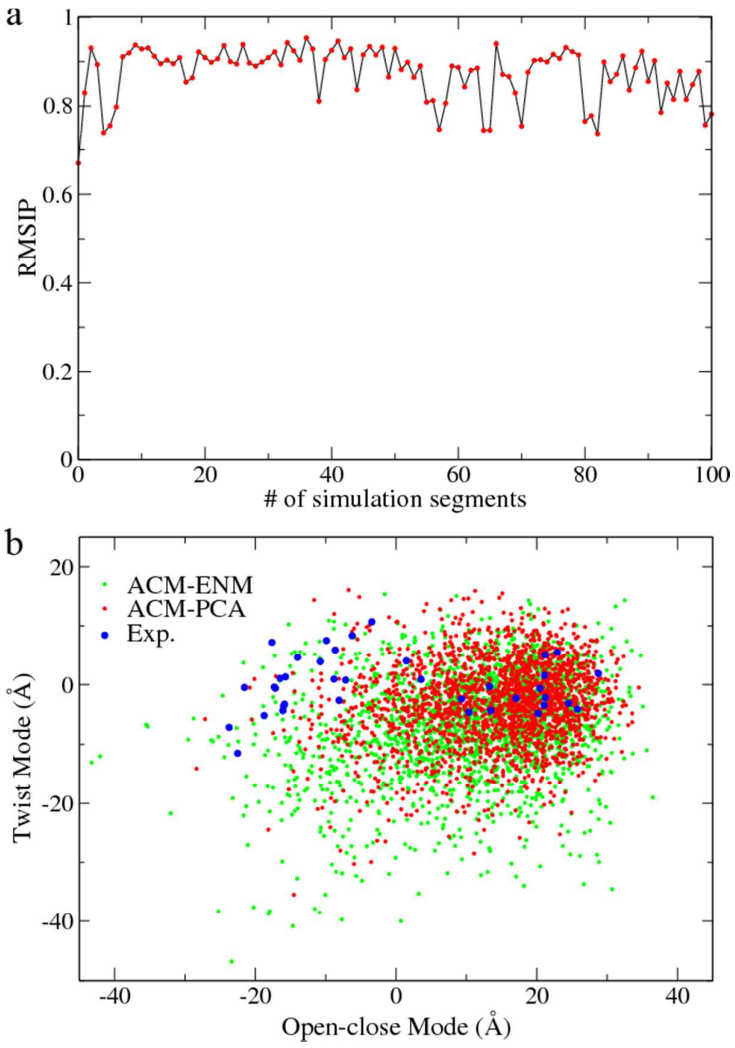

Figure S7. (a) RMSIP values between the PCA modes calculated from each simulation segment (200 ps) and the ENM modes from the final structure of this trajectory. (b) Projection of the ACM-ENM (green) and ACM-PCA trajectory (red) onto the 2D essential subspace, respectively.

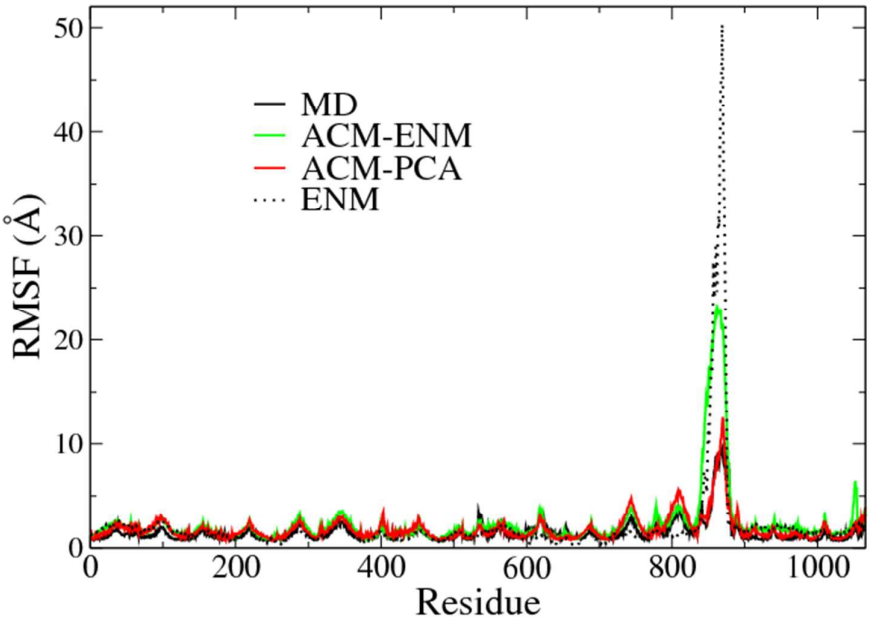

Figure S8. RMSF of the COM of residues in the vinculin, which were calculated from the MD (black solid line), ACM-ENM (green solid line), and ACM-PCA (red solid line) simulation, respectively. These RMSF values predicted from ENM (Eqn. S5) is also shown (black dash line), which indicate there is significant tip effect in the PR loop.

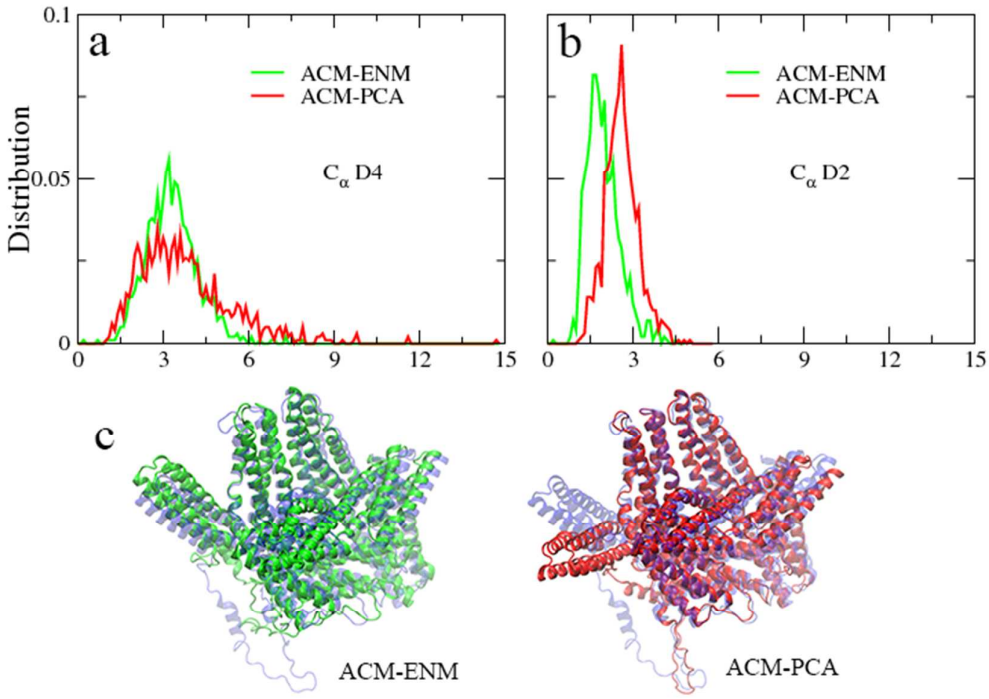

Figure S9. Distributions of RMSD in the ACM-ENM (green solid line) and ACM-PCA (red solid line) simulation, respectively. The RMSD values were calculated using the C_α atoms of (a) D4, and (b) D2. All the conformations were superimposed to the starting structure using the C_α atoms of D1+D3. (c) Domain motions of the vinculin in ACM-ENM and ACM-PCA. In each simulation, the starting structure is colored by blue and the structure with the largest RMSD of D4 to D1+D3 is colored by green in ACM-ENM, and red in ACM-PCA, respectively.

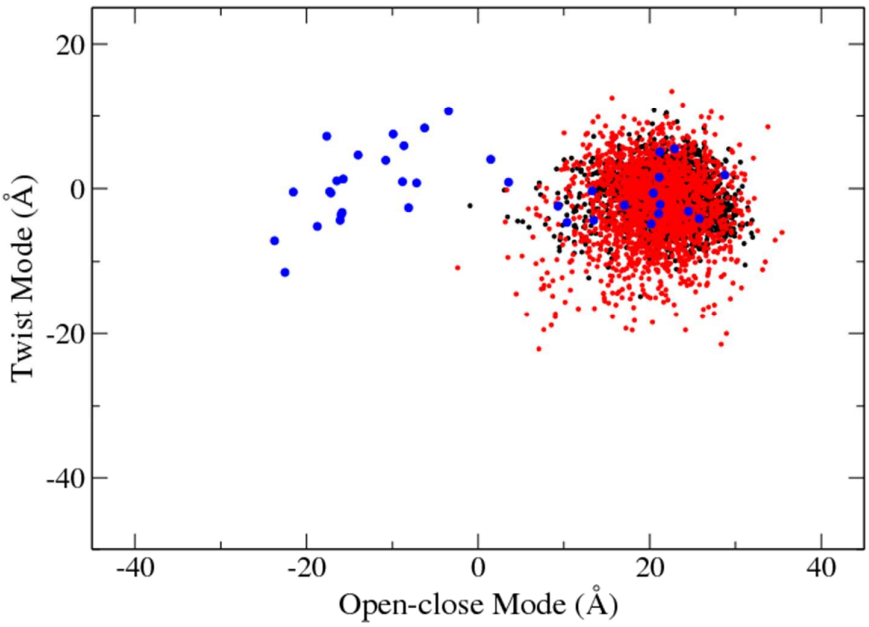

Figure S10. The ACM-PCA simulation without updating the PCA modes. PCA was performed on the 200 ns MD trajectory, and then the first three modes were used to run a 20 ns ACM-PCA simulation without further updating. All the conformations in the trajectory are projected on to the 2D essential subspace. The MD trajectory is also projected for comparison (black).