Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supporting Information

How Low-Resolution Structural Data Predict the Conformational Changes of a Protein: A Study on Data-Driven Molecular Dynamics Simulations

Ryuhei Harada*† and Yasuteru Shigeta*†

Center for Computational Sciences, University of Tsukuba, 1-1-1 Tennodai, Ibaraki 305-8577, Japan

Markov state model construction

After the hybrid conformational sampling, the trajectories of OFLOOD (25 cycles) were utilized to construct MSM. To construct MSM, a MSM builder (EMMA)[1] was adopted. The construction of an appropriate MSM was done within the two steps. First the trajectories were clustered and then discretized to define microstates. For the clustering, the k-means clustering algorithm was selected for the computation of a total number of 50 clusters (microstates). Then each frame of the trajectories was assign to the closet cluster. To determine an approximate lag time, the implied times t_i were estimated as follows:

$$t_i(\tau) = -\frac{\tau}{\ln \lambda_i(\tau)}, \quad (1)$$

where τ is the lag time, and denotes the *i*th slowest implied time scale determined from the *i*th largest eigenvalue of a transition matrix $\boldsymbol{T}$. When $t_i(\tau)$ reaches an approximately constant value with increasing lag time τ, then it means that target system starts to satisfy the Markov assumption. Figure S1 shows the implied time scale as a function of τ. Judging from these profiles, the implied time scale curves reach a constant when τ goes beyond around 140 ps. Therefore, τ of 140 ps was chosen for the present demonstration. In terms of the lag time, EMMA counted the transitions among the microstates and estimate the maximum likelihood transition matrix $\boldsymbol{T}$ under the constrained detailed balance. The constructed transition matrix had the dimension 50×50 according to the number of microstates.

After constructing MSM, the resulting model was validated based on the Chapman-Kolmogorov test using the Chapman-Kolmogorov equation:

$$\boldsymbol{T}(n\tau) \approx \boldsymbol{T}(\tau)^n, \quad (2)$$

where *n* is an integer number of steps. Theoretically, the Chapman-Kolmogorov test is usually conducted by comparing the probability of remaining in a selected state at increasing time steps of the MD trajectories with that of MSM predictions. This test was performed using a tool of EMMA. In this test, the probability decay of major metastable states predicted by MSM was compared to those estimated from the original simulations

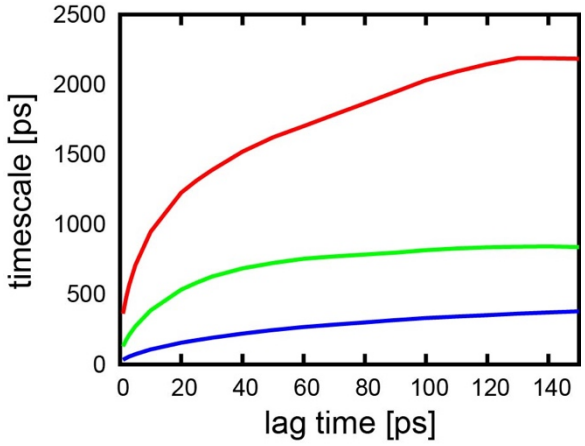
(see Figure S2). Judging from these profiles, the estimated curve of the probabilities decay matched well with the predicted one, indicating that the resulting model is Markovian.

After validating the constructed MSM, the FEL of LAO protein was calculated. To estimate the FEL, one has to obtain an equilibrium distribution from the transition matrix, $\boldsymbol{T}$. Generally, it is easy to estimate the equilibrium distribution $\boldsymbol{\pi}$ when the target process is the Markov process and equilibrium enough, i.e. the transition among the microstates are described as $\boldsymbol{\pi} = \boldsymbol{T}\boldsymbol{\pi}$, i.e. $\boldsymbol{\pi}$ corresponds to the normalized equilibrium distribution constructed from a set of distributions, $\{\pi_i\}$ ($\sum_i \pi_i = 1$). Therefore, $\boldsymbol{\pi}$ was obtained from one of the eigenvector of $\boldsymbol{T}$. Finally, FEL was calculated as follows:

$$F_i = -k_B T \ln \frac{\pi_i}{\max_j \pi_j} \ (i = 1, 2, \ldots, 50), \tag{3}$$

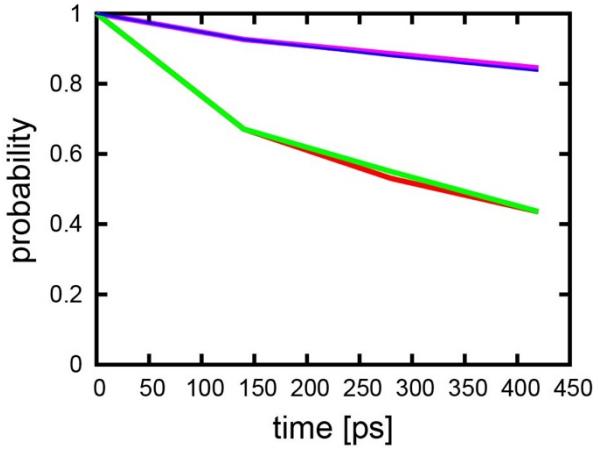
where the origin of FEL was selected as the maximum value of $\boldsymbol{\pi}$. Actually, the apo-type FEL of LAO protein (without the ligand) was projected onto the conformational subspace spanned by $RMSD_{open}$ and $RMSD_{closed}$ (Figure 7).

Figure S1



Profiles of the implied times as a function of the lag time τ for the top three eigenvalues of the transition matrix ***T*** presumed with EMMA, corresponding to the red, green, and blue lines.

Figure S2



Chapman-Kolomogrov test of the MSM conducted at τ = 140 ps with 50 microstate defined by the k-means clustering for (two) major metastable states, corresponding to the probability decay and the predicted one: (1) the pair of the red/green and (2) magenta/blue lines.

References

1. M. Senne, B. Trendelkamp-Schroer, A. S. J. S. Mey, C. Schutte and F. Noe, *J. Chem. Theory Comput.*, 2012, **8**, 2223-2238.