Recent Extensions and Applications of Parallel Cascade Selection Molecular Dynamics Simulations

Ryuhei HARADA and Yasuteru SHIGETA

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

1. Introduction

Proteins have incredibly complex structures and functions related to its structural dynamics, which are fundamental to various biological phenomena occurring *in vivo* such as molecular recognition, transport processes, enzymatic reactions, *etc*. Although more than 130,000 structures are currently stored in the Protein Data Bank (PDB), the direct measurement of structural changes, which are essential for a protein function, was quite limited in spite of the recent advances in single-molecule experiments. Alternatively, conventional molecular dynamics (CMD) are ways to reproduce biological phenomena *in silico*. However, it is often difficult for CMD to keep track with real biological phenomena owing to its accessible timescale. Thus, in order to theoretically observe large-scale structural changes such as protein folding and domain motion, extremely long-time dynamics [1-3], multi-canonical method [4-5], replica exchange method [6], metadynamics [7-9], temperature-accelerated MD method [10], *etc* were applied to overcome the timescale issue. However, they still need know-how for the individual problem and huge computer resources. Therefore, it is desirable to develop easier and automatic methods for transition pathway sampling of complicated systems such as proteins.

As a more straightforward and faster sampling method, we have proposed an efficient sampling method consisting of (1) an initial structure selection with high possibility of inducing structural changes and (2) structural resampling of initial structure based on short time MD. The key to this method is to select plausible initial structures from past trajectories. Once a transition path, which connects one stable structure to the other, is obtained, precise free energy analyses can be performed immediately. Therefore, this method might be more efficient than the other conformational sampling methods.

In this short review, the calculation schemes and several examples of structural changes of proteins are outlined (for details see the review paper [11]). In the following section, we explain the methodologies of our approach. Then, numerical results are discussed in Sec. 3. Concluding remarks are given in Sec. 4.

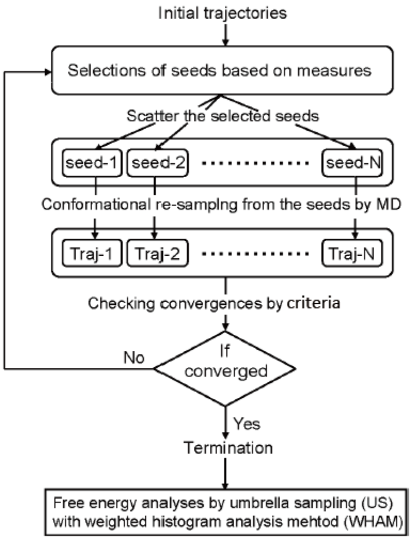
Initial trajectories
Selections of seeds based on measures
Scatter the selected seeds
seed-1
seed-2
seed-N
Conformational re-samplng from the seeds by MD
Traj-1
Traj-2
Traj-N
Checking convergences by criteria
No
If converged
Yes
Termination
Free energy analyses by umbrella sampling (US) with weighted histogram analysis mehtod (WHAM)

Figure 1. A flowchart of our conformational sampling scheme

2. Methodology

2.1. Basic Idea of Parallel Cascaed Selection Molecular Dymamics (PaCS-MD)

The method we have proposed is rather simple. The flowchart of the algorithm is shown in **Figure 1**. The calculation procedure of the structural sampling method is briefly described below.

I. Executing MD simulations from several selected initial structures ($n_{\mathrm{initial}} = N$) for a relatively short time with a canonical ensemble (*NVT* or *NPT*).

II. Ranking snapshots of each MD trajectory based on a predetermined rule (selection rule) for some measures.

III. Selecting the n_{initial} structures of the next cycle, where snapshots with higher rank are preferentially selected.

IV. Regenerating the initial velocities at the target temperature according to the Maxwell-Boltzmann distribution.

V. Repeating the cycles of (I) - (IV) until the distribution function along certain reaction coordinates does not change any more. Otherwise, the cycle ends when some of structures (or data) are sufficiently close to the target structure (or data).

VI. Applying the umbrella sampling (US) method [12-13] followed by the weighted histogram analysis method (WHAM) [14-15] to obtain free energy landscape projected onto reaction coordinates by referring a reactive trajectory obtained above. Markov State Model is also utilized to estimate free energy surfaces in terms of trajectories sampled by PaCS-MD.

The original methodology, i.e. PaCS-MD [16], ranks snapshots based on the similarity between the trajectory obtained from the short-time MD and the target structure (for example, root-mean square deviation (RMSD) between them). After determining the ranks, one selects the n_{initial} (typically 10 – 100) structures with higher rank as the initial structures of the next cycle and performs n_{initial} different short-time (typically 100 ps) MD simulations independently. By repeating the series of cycles, PaCS-MD generates closer structures to the target structure than those found in the previous cycle and occasionally induces structural transitions without using external perturbations.

2.2 Variants of PaCS-MD without using target structures

The disadvantage of the original PaCS-MD is to require the target state. To remove the drawback, we have developed several variants discussed below in detail. A conceptual difference from PaCS-MD is to use a different selection rule in each method, which has each advantage and disadvantage compared to PaCS-MD. Unlike PaCS-MD, its variants, i.e. fluctuation flooding method (FFM) [22], outlier flooding (OFLOOD) [24], structural dissimilarity sampling (SDS) [29] methods mentioned below, do not require a target (product) structure *a priori*, while these method take more cycles to find transition pathway than PaCS-MD does.

Dynamics of proteins involved in functional expression are often anisotropic, and certain vibrational modes with high anharmonicity are dominant. We consider that the anisotropy can be extracted by principal component analysis (PCA). We have developed FFM that efficiently induces structure transitions by assuming that protein structure with high structural fluctuation transits with high probability with structural re-sampling [22]. The basic concept of FFM is that the principal modes with large eigenvalues are chosen as reaction coordinates of a large amplitude motion and used for estimating measures. The snapshots with the maximum and minimum inner product values for the *j*-th principal component coordinates (PC*j*) and/or a multi-dimensinal PC vector space, which is obtained by PCA, are selected as the initial structures at the next cycle.

Metastable states exist in a high-dimensional subspace, where high density distributions appear during MD simulations and are detected as clusters, and transitions among different metastable states occur with large structural changes. Since the sparse distribution exists among clusters (sometimes between two clusters), a structure change is induced by intensively selecting structures with lower density distribution as the initials. Sparse distributions that do not belong to clusters are referred as "outliers" and can be detected using a hierarchical clustering method such as FlexDice [23]. Structural resampling intensively from the outliers of distribution is called as "OFLOOD” method [24], which promotes structural transformation efficiently in addition to its variants [24-28].

SDS, which realizes efficient structural sampling by repeating structural resampling, selects the initial structures so that the structural correlation with a structure (starting structure, mean structure, and so on) becomes as small as possible in the current cycle. This method is an effective technique when the target structure is unknown and a dissociation process of a molecular copmlex. For details of the calculation procedure, see the previous works [29-30].

2.3. Numerical Details

Initial structural data were taken from the PDB as the starting structures of PaCS-MD and its variants. After solvation with SPC/E or TIP3P water models and neutralization with counter ions, CMD simulations were performed under three-dimensional periodic boundary conditions using the Amber force field [31]. Water molecules were treated as rigid bodies with the SETTLE algorithm [32], while chemical bonds of the proteins were treated as rigid bodies with the LINCS algorithm [33]. To model the equilibrated systems, *NVT* simulations were first performed and followed by *NPT* simulations. *NVT* and *NPT* simulations were conducted with the modified Berendsen thermostat at 300 K [34] and the Parrinello-Rahman method at 1 bar and 300 K [35-36]. The equations of motion were integrated by the leapfrog method. The time step length was set to 2 fs. The cutoff value for Coulomb and van der Walls interactions was set to 10 Å. During conformational resampling, trajectories were recorded every 1 ps. For the PaCS-MD and its variant simulations, 10 or 100 different initial structures (n_{initial}) were simultaneously used for the efficient conformational sampling depending on the biological systems At every cycle, 100 ps short-time MD simulation with renewed velocity under *NVT* ensemble ($T = 300$ K) were launched until sufficient conformational sampling was accomplished. For the free energy analysis, 200 reference structures were randomly selected from the trajectories of FFM and the umbrella sampling for each reference structure was performed for 1ns and followed by WHAM to estimate the free energy landscape of T4L. All MD simulations were performed with the GPU version of the Gromacs 5.0.7 package [37].

3. Results and Discussion

3.1 Domain motion of T4L by FFM

As an example of FFM, we here explain results of an open-closed structure transition of T4L. The open structure of T4L (wild type) was chosen as a starting structure. In order to induce structural transition, we selected the first and second principal coordinates (PC_1 and PC_2) as the reaction coordinates. In order to obtain PCs, CMD (10 ns) from the open structure was performed *a priori*, then PCA was performed using the 10-ns trajectory. According to the accumulated contribution of principal modes (PMs), the open-closed conformational transition can be sufficiently described because

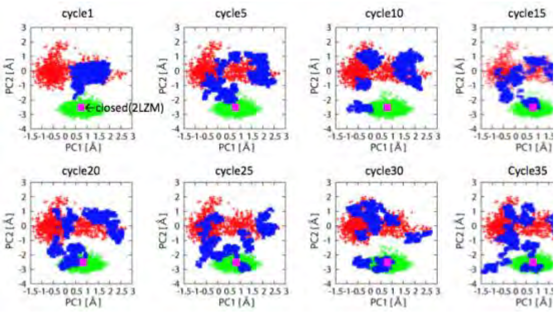

Figure 2: Conformational transition pathways from the open to the closed states of T4 lysozyme reproduced by FFM.

the top two PMs accounted for 80% or even more of the overall protein structure fluctuation. Here, we show the structural transition process from the open to closed structures of T4L reproduced by FFM [16]. **Figure 2** shows the accumulated distributions of trajectories projected onto a subspace spanned by PC_1 and PC_2. For comparison, we also depicted red points (green points) which show the projection of a conventional MD trajectory (10 ns) from the open structure (the closed structure). The projected points by FFM (blue points) expand to the periphery and reaches the closed structure (magenta) within about 15 cycles (several tens of ns). After 15th cycles, the accumulated trajectories travels all over regions in the subspace. As a comparative calculation, extremely long-time CMD simulation (1 μs) was started from the open structure. However, no open-close conformational transition was observed. After FFM, we estimated FEL onto the subspace as illustrated in **Fig. 3**, whose cost was 200 ns, indicating the efficient structure search and FEL analyses of FFM. According to this result, the free energy barrier and transition states are clearly found between two states.

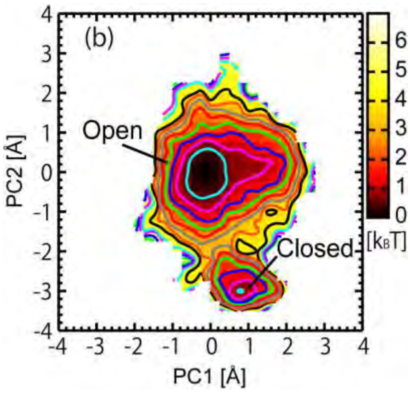

Figure 3: Conformational transition pathways from the open to the closed states of T4 lysozyme reproduced by FFM.

3.2 Protein folding pathway of Villin by OFLOOD method

As an application example of the OFLOOD method, we here show the protein folding process of the small protein Villin (35 residues). Partial RMSDs, helix 1 - helix 2 (segment A) and helix 2 - helix 3 (segment B), measured from the native structure were used as the reaction coordinates, which had been defined in the previous study [32]. After modeling the amino acid chain, structural sampling by the OFLOOD method with 100 seeds was repeated for 20 cycles, where the generalized Born and surface area (GB/SA) solvation model was adopted for the solvent around the Villin.

The projection of the trajectories generated by the OFLOOD method is shown in **Figure 4**. It is easily found that the outliers (black points) were located the edge of the distribution at each cycle and became broaden as cycles went by. It took about 10 cycles to reach the native structure of Villin highlited by a cross in **Figure 4**. The minimum C_α RMSD measured from the X-ray crystal structure after the end of 20 cycles was 0.60 Å, indicating that natural structure can be sampled by the OFLOOD method accurately. Furthermore, we also extracted a minor pathway (see right bottom of **Figure 4**) that could not be observed in the

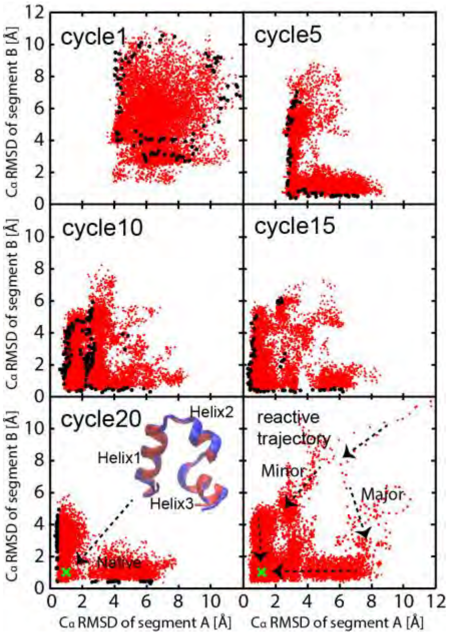

Figure 4. Major and minor folding pathways of Villin reproduced by OFLOOD method

previous study [4]. Concerning to the computational efficiency, the cumulative computation time to sample the native structure (C_α RMSD <1.0 Å) was 135.6 ns. We could extract the folding path very efficiently compared to the replica exchange MD (8 μs) [4]. Since computational cost required in our calculations was ns-order and the time-scale of the protein folding process is μs-order, the OFLOOD method is quite efficient for finding protein folding pathways.

3.3 Open-to-Closed structure transition pathway searches by SDS

To show the conformational sampling efficiency, SDS was applied to structural transition between two different states (open and closed states) of maltodextrin binding protein (MBP) in explicit water. MBP is a protein consisting of 370 residues and induces large-amplitude domain motions for ligand binding. In this example, apo-type simulations were considered. CMD simulation may be useful for reproducing the large-amplitude (open-closed) domain motions of MBP. However, we have not detected the structural transition at all, even if a long-time (1 μsec) simulation was perform starting from the open state. Instead we have performed SDS simulations for 50 cycles starting both from the apo-type closed and open states. As a result of the demonstration starting from the open (closed) structures, the minimum values of Cα RMSD measured from the X-ray crystal structure of closed (open) forms were 0.78 Å (0.89 Å) as illustrated in **Fig. 5**. The closest snapshots to the X-ray crystal structures sampled by the SDS simulations are overlapped

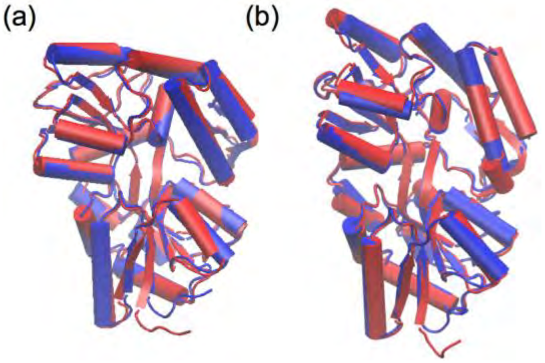

Figure 5: (a) Snapshot with the minimum value of RMSD (0.78 Å) measured from the closed state of MBP (red), which was sampled by SDS during the 50 cycles starting from the open state of MBP. (b) Snapshot with the minimum value of RMSD (0.89 Å) measured from the open state of MBP (red). Each snapshot is superimposed with the X-ray crystal structure (blue).

in **Figure 5(a)** and **5(b)**, indicating that SDS also induces the open-to-closed and its inverse conformational transitions quite efficiently and gave accurate structures without knowledge of the well-defined target structures.

4. Conclusion

In this review, we outline the PaCS-MD and its variants and show examples of their applications. If the reactant and the protein structure of the product are known, PaCS-MD can be applied to extract the path connecting the two end-point structures. On the other hand, FFM, OFLOOD, and SDS are applicable to extract transitional pathways starting from the protein structure of a given reactant without knowldge of any product. For example, FFM started with an open structure and derives an open-closed domain motion of T4L using ns-order simulation. OFLOOD also presumed Villin's folding pathways using ns-order simulations starting with fully extended structures. SDS generates both open-to-closed and closed-to-open structural transitions of MBP within 500 ns-order simulations. These types of methods may be convenient for predicting metastatic pathways or starting from a given reactant without knowledge of the reactants and finding the local/global energy minimum state of proteins , Which is one of the advantages of three-dimensional sampling.

Acknowledgement

This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant-in-Aid for Young Scientists (A) (No. JP16H06164), and by Grant-in-Aid for Scientific Research of the Innovative Areas (Nos. JP26102525, JP 26107004, and JP 26105012) from JSPS. The computations were performed using the RIKEN Integrated Cluster of Clusters (RICC), Super Computer Center of Institute of Solid State Physics, The University of Tokyo and COMA / HP-PACS at the CCS, University of Tsukuba.

References

[1] D.E. Shaw, *et al*., *Commun. Acm.*, **51** (2008) 91.

[2] D.E. Shaw, *et al*., *Science*, **330** (2010) 34.

[3] K. Lindorff-Larsen, S. Piana, R.O. Dror, D.E. Shaw, *Science*, **334** (2011) 517

[4] N. Nakajima, H. Nakamura, A. Kidera, *J. Phys. Chem. B*, **101** (1997) 817.

[5] U.H.E. Hansmann, Y. Okamoto, F. Eisenmenger, *Chem. Phys. Lett.* **259** (1996) 321.

[6] Y. Sugita, Y. Okamoto, *Chem. Phys. Lett.*, **314** (1999) 141.

[7] A. Laio, M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.* **99** (2002) 12562.

[8] A. Laio, F.L. Gervasio, *Rep. Prog. Phys.* **71** (2008) 126601.

[9] P. Raiteri, A. Laio, F.L. Gervasio, C. Micheletti, M. Parrinello, *J. Phys. Chem.*

B, **110** (2006) 3533.

[10] C.F. Abrams, E. Vanden-Eijnden, *Proc. Natl. Acad. Sci. U. S. A.* **107** (2010) 4961.

[11] R. Harada, Y. Takano, T. Baba, Y. Shigeta, *Phys. Chem. Chem. Phys.*, **17** (2015) 6155.

[12] G.M. Torrie, J.P. Valleau, *J. Comput. Phys.*, **23** (1977) 187.

[13] G.M. Torrie, J.P. Valleau, *J. Chem. Phys.*, **66** (1977)1402.

[14] A.M. Ferrenberg, R.H. Swendsen, *Phys. Rev. Lett.*, **63** (1989) 1195.

[15] M. Souaille, B. Roux, *Comput. Phys. Commun.*, **135** (2001) 40.

[16] R. Harada, A. Kitao, *J. Chem. Phys.*, **139** (2013) 035103.

[17] R. Harada, A. Kitao, *J. Chem. Theory Comput.*, **11** (2015) 5493.

[18] R. Harada, Y. Shigeta, *J. Comput. Chem.*, **38**, 2671, (2017).

[19] R. Harada, Y. Shigeta, *Bull. Chem. Soc. Jpn.*, **90** (2017) 1236.

[20] R. Harada, Y. Shigeta, *Chem. Lett.*, **46** (2017) 862.

[21] R. Harada, Y. Shigeta, *Mol. Simul.*, **44** (2017) 206.

[22] R. Harada, Y. Takano, Y. Shigeta, *J. Chem. Phys.*, **140** (2014) 125103.

[23] T. Nkamura, Y. Kamidoi, N. Wakabayashi, N. Yoshida, *IPSJ J. Database*, **46** (2005) 40.

[24] R. Harada, T. Nakamura, Y. Takano, Y. Shigeta, *J. Comput. Chem.*, **36** (2015) 97.

[25] R. Harada, T. Nakamura, Y. Shigeta, *Chem. Phys. Lett.*, **639** (2015) 269.

[26] R. Harada, T. Nakamura, Y. Shigeta, *J. Comput. Chem.*, **37** (2016) 724.

[27] R. Harada, T. Nakamura, Y. Shigeta, *Bull. Chem. Soc. Jpn.*, **89** (2016) 1361.

[28] R. Harada, Y. Takano, Y. Shigeta, *J. Comput. Chem.*, **38** (2017) 790.

[29] R. Harada, Y. Shigeta, *J. Comput. Chem.*, **38** (2017) 1921.

[30] R. Harada, Y. Shigeta, *Phys. Chem. Chem. Phys. acceped for publication* (2018).

[31] Y. Duan, *et al.*, *Comput. Chem.* **24** (2003) 1999.

[32] S.Miyamoto, P.A. Kollman, *J. Comput. Chem.* **13** (1992) 952.

[33] B. Hess, H. Bekker, H.J. Berendsen, J.G.E. Fraaije, *J. Comput. Chem.* **18** (1997) 1463.

[34] G. Bussi, D. Donadio, M. Parrinello, *J. Chem. Phys.* **126** (2007) 014101.

[35] S. Nose, M.L. Klein, *Mol. Phys.* **50** (1983) 1055.

[36] M. Parrinello, A. Rahman, *J. Appl. Phys.* **52** (1981) 7182.

[37] M.J. Abraham, D. van der Spoel, E. Lindahl, B. Hess, and the GROMACS development team, *GROMACS User Manual version 5.0.7*, www.gromacs.org (2015).