Simple, yet powerful methodologies for conformational sampling of proteins

Journal:	*Physical Chemistry Chemical Physics*
Manuscript ID:	CP-PER-11-2014-005262.R1
Article Type:	Perspective
Date Submitted by the Author:	20-Jan-2015
Complete List of Authors:	Harada, Ryuhei; Tsukuba University, Center for Computational Sciences Takano, Yu; Osaka University, Institute for Protein Research Baba, Takeshi; Osaka University, Department of Materials Engineering Science, Graduate School of Engineering Science Shigeta, Yasuteru; University of Tsukuba, Physics

Simple, Yet Powerful Methodologies for Conformational Sampling of Proteins

[a,b] Ryuhei Harada, [b,c] Yu Takano, [d] Takeshi Baba, and [a, b, e] Yasuteru Shigeta

[a] Center for Computational Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8571, Japan.
[b] JST, CREST, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan.
[c] Institute of Protein Research, Osaka University, Suita, Osaka 565-0871, Japan.
[d] Department of Materials Engineering Science, Osaka University, Toyonaka, Osaka 560-8531, Japan.
[e] Department of Physics, Graduate School of Pure and Applied Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8571, Japan.

The corresponding authors: R. Harada and Y. Shigeta

Abstract

Several biological functions, such as molecular recognition, enzyme catalysis, signal transduction, allosteric regulation, and protein folding are strongly related to conformational transitions of proteins. These conformational transitions are generally induced as slow dynamics upon collective motions, including biologically relevant large-amplitude fluctuations of proteins. Although molecular dynamics (MD) simulation has become a powerful tool for extracting conformational transitions of proteins, it might still be difficult to reach time scales of the biological functions because the accessible time scales of MD simulations are far from biological time scales, even if brute-force conventional MD (CMD) simulations using massively parallel computers are employed. Thus, it is desirable to develop efficient methods to achieve canonical ensembles with low computational costs.

From this perspective, we review several enhanced conformational sampling techniques of biomolecules developed by us. In our methods, multiple independent short-time MD simulations are employed instead of single brute-force long-time CMD simulations. Our basic strategy is as follows: (i) selection of initial seeds (initial structures) for the conformational sampling in restarting MD simulations. Here, the seeds should be selected as candidates with high potential to transit. (ii) Resampling from the selected seeds by initializing velocities in restarting short-time MD simulations. A cycle of these simple protocols might drastically promote the conformational transitions of biomolecules. (iii) Once reactive trajectories extracted from the cycles of short-time MD simulations are obtained, a free energy profile is evaluated by means of umbrella sampling (US) techniques with the weighted histogram analysis method (WHAM) as a post-processing technique.

For the selection of the initial seeds, we proposed four different choices: (1) Parallel CaScade Molecular Dynamics (PaCS-MD), (2) Fluctuation Flooding Method (FFM), (3) Outlier FLOODing (OFLOOD) method, and (4) TaBoo SeArch (TBSA) method. We demonstrate applications of our methods to several biological systems, such as domain motions of proteins with large-amplitude fluctuations, conformational transitions upon ligand binding, and protein folding/refolding to native structures of proteins. Finally, we show the conformational sampling efficiencies of our methods compared with those by CMD simulations and other previously developed enhanced conformational sampling methods.

Abbreviations: Molecular Dynamics (MD), conventional MD (CMD), Weighted Histogram Analysis Method (WHAM), Parallel CaScade Molecular Dynamics (PaCS-MD), Fluctuation Flooding Method (FFM), Outlier FLOODing (OFLOOD) method, TaBoo SeArch (TBSA) method, Transition Path Sampling (TPS) method, Reaction Coordinate (RC), Minimum Free Energy Pathway (MFEP), Collective Variable (CV), Targeted MD (TMD), Steered MD (SMD), Meta Dynamics (MetaD), Replica Exchange MD (REMD), Multicanonical MD (McMD), Density of States (DOS), Degrees of Freedom (DOF), Coarse-Grained (CG), All-Atom (AA), Hamiltonian REMD (HREMD), Free Energy Landscape (FEL), MultiScale Free Energy Landscape (MSFEL) method, MultiScale Essential Sampling (MSES), Umbrella Sampling (US), Principal Component Analysis (PCA), Principal Mode (PM), Principal Coordinate (PC), Root Mean Square Deviation (RMSD), Nuclear Magnetic Resonance (NMR), T4 Lysozyme (T4L), Glutamine-Binding Protein (GlnBP), Lysine-, Arginine-, Ornithine-binding protein (LAO), Maltose-Binding Protein (MBP), Full Correlation Analysis (FCA), Independent Component Analysis (ICA), Markov State Model (MSM), Logarithmic Mean-Force Dynamics (LogMFD), Triplet–Triplet Energy Transfer (TTET) experiment.

1. Introduction

Biological functions are determined based on both structural and dynamical aspects of biomolecules. From the structural aspect, for instance, X-ray crystallography and nuclear magnetic resonance (NMR) experiments provide structural information of biomolecules as reduced low-dimensional data such as their electron densities or distances between hydrogen atoms. Based on the reduced low-dimensional data, three-dimensional structures of the biomolecules are reconstructed. Furthermore, structural biology provides coarse-grained information on the biological functions based on “sequence-to-structure” relationships. To obtain more detailed dynamical information on the biological functions, molecular dynamics (MD) simulations on the basis of experimental information might directly keep track of time series of atomic coordinates of biomolecules.[1, 2] In this sense, MD simulation is indispensable for extracting the dynamics of biomolecules relevant to the biological functions. However, accessible time scales of conventional MD (CMD) simulations with canonical ensembles are limited to several hundreds of nanoseconds by using laboratory-level computer resources, which are often very far from the time scales of biological functions. In addition, these biological functions are induced as rare events on the micro- to millisecond time scales as stochastic processes. Even if state-of-the-art computational resources are employed, it is difficult to reach the time scales of relevant biological functions, leading to the quasi-ergodicity problem because of insufficient conformational sampling of the rare events stochastically induced in long-time MD simulations. Thus, one cannot predict when the biological processes will occur. There exist two different routes to remedy this problem. One is to extend the accessible time scales of the MD simulations towards those of biological functions through the development of algorithms and computer architectures for accelerating MD simulations themselves. The other is the development of enhanced conformational sampling methods for extracting biologically relevant rare events stochastically induced in the long-time MD simulations, which are important for obtaining reliable canonical ensembles.

For the former, the Shaw group has recently achieved significant advances in the computational efficiency of MD simulations through their development of a specialized machine called “Anton”.[3, 4] In Anton, the entire MD simulations are performed by their custom-designed computer architecture specialized for MD simulations. His group has performed all-atom MD simulations of some typically fast-folding miniproteins explicitly including water for micro- to millisecond time scales and found their reversible folding and unfolding processes. Quite recently, his group successfully simulated the folding/refolding processes of Ubiquitin (76 residues) in an explicit solvent environment through millisecond-order MD simulations.[5] Their results stimulate us to expect that these brute-force MD simulations might unveil important biological processes at an atomic level. Along with the specialized hardware, general-purpose graphics processing units (GPUs) have been frequently used with MD simulation programs because a GPU is cheaper than the specialized hardware noted above and can be applicable to various applications. Many important algorithms and models have been implemented on GPU-based MD programs,[6–11] in which several orders of magnitude over the CPU implementations have been achieved. For example, myPresto/psygene-G,[7] which is a recently developed GPU-based MD program, has accelerated calculations of the electrostatic interaction using zero dipole summation methods,[8-10] which is a suitable algorithm for the GPU architecture because it avoids the Ewald summation. Because the current popular MD programs, including AMBER,[11, 12] Gromacs,[13] NAMD,[14] and OpenMM[15] have been implemented on GPUs, nowadays one can use them for long-time simulation of large biomolecules. However, even if one uses these machines, one cannot escape from the stochastic problem of the rare events, *i.e.*, reproducibility of these events in another long-time simulation with different initial conditions.

For the latter, as a complementary approach to the acceleration of MD simulations, several efficient conformational sampling methods had been extensively developed before the emergence of Anton. In these methods, external perturbations such as biased forces or modified potentials are artificially imposed to promote the conformational transitions. In general, these methods can be categorized into two types. The first type is to sample transition pathways under a condition that two end-point structures, *i.e.*, a set of reactant and product, are given *a priori*. For instance, the Transition Path Sampling (TPS)[16-18] method provides ensembles of transition pathways connecting the reactant to the product. TPS refines an initial transition pathway through a random walk in transition path space without specifying a set of reaction co-ordinates (RCs). However, convergence of transition paths is so sensitive to an initial guess that the initial guess should be as reasonable as possible. Nevertheless, it is difficult to prepare an appropriate initial guess *a priori*. As another transition path search method, String Method[19-23] is powerful in obtaining a relaxed transition pathway on a Minimum Free Energy Pathway (MFEP).[24] In the String Method, coarse-grained collective variables (CVs) to describe the conformational transitions should be specified *a priori*. Employing CVs, multiple images on an initial transition pathway are gradually relaxed to an optimal pathway on MFEP under weak restraints with the neighbouring images. As a biased dynamics, Targeted MD (TMD)[25] or Steered MD (SMD)[26, 27] provide sets of candidates of transition pathways connecting the reactant to the product by imposing external restraints or forces with respect to the product. If the transition pathways obtained by these methods have some hysteresis, the imposed restraints and forces might be inappropriate. However, validation of the obtained transition pathways is quite difficult because of dependence of the external restraints or forces on the transition pathways. Therefore, it is desirable to double-check one's data after TMD or SMD.

Meta Dynamics (MetaD)[28-30] is another enhanced conformational sampling method based on external bias potentials and has been applied to several complicated biological reactions.[31-35] An assumption in MetaD is that biological reactions of systems can be described by a few CVs. Once CVs are specified, positive history-dependent Gaussian potentials described by the CVs should be added to the potential energy to enhance the conformational sampling of the systems. As a result of the external biases, conformational flooding through a positive Gaussian potential discourages the system from returning to the previously sampled structures. These positive Gaussian potentials are imposed until the trajectories have sufficiently explored the energy landscapes. One of the advantages of MetaD is that accumulated history-dependent potentials asymptotically converge to the free energy surface spanned by the CVs. However, for the termination of MetaD, it is generally difficult to judge whether the conformational flooding has converged or not. Furthermore, the conformational sampling efficiency tends to depend strongly on the specified CVs, and how to specify the CVs is a non-trivial issue in any application.

The second type is to find efficiently global minimum and metastable structures of biomolecules with low free energy. Because there are numerous configurations even for small proteins in their configuration space, the enumeration of biologically relevant structures from all possible structures is quite difficult. For instance, in folding simulations of proteins, the ultimate goal is to find the global minimum on a folding funnel starting from a fully stretched structure. To find the global minimum in the folding funnel, Replica Exchange MD (REMD),[36] Multicanonical MD (McMD)[37, 38] simulations and their variants,[39-42] which are referred to as the generalized ensemble method,[43] have been widely used. With these generalized ensemble methods, conformational sampling with the biased weight can be reweighted to generate non-biased canonical ensembles at the target temperature, therefore it is convenient for energetic analyses of biomolecules such as structural stabilities among metastable states. In REMD, multiple replicas with different temperatures are prepared and parallel MD simulations for the replicas at each temperature are performed. During the parallel MD simulations, the replicas with neighbouring temperatures are exchanged so as to satisfy a detailed-balance condition. Because of the exchanging from low to high temperatures, systems might escape from local minima because they have sufficient kinetic energy to cross the energy barriers among metastable states. Apart from the Boltzmann weight sampling in REMD, a non-Boltzmann weight sampling is employed to enhance the conformational sampling in McMD. To promote the conformational transitions, McMD uses modified potential functions, which are evaluated from the inverse density of states (DOS), so that a random walk in the flattened potential energy space leads to efficient conformational sampling. However, as it often takes a long time in the McMD simulations, the DOS should be accurately estimated before the production run of McMD. REMD and McMD have common difficulties in applications to large systems because preparation of initial parameter settings becomes quite tedious as systems become large and complicated. For instance, the number of replicas necessary for sufficient conformational sampling exponentially increases to maintain a high exchange ratio between the replicas in REMD. In McMD, DOSs to be estimated for broad conformational sampling on the potential energy space also exponentially increase with system size. That is why the applications of these methods are currently limited to fast-folding miniproteins with less than 50 residues. Moreover, it is generally difficult to restore details of dynamics because the accelerated dynamics track motions on modified potentials, which are different from the real potentials. In other words, the time series of detailed atomic trajectories are discarded in compensation for increasing the conformational sampling efficiencies.

As another type of efficient conformational sampling method, several "multiscale" simulation methods have been proposed recently. Here, "multiscale" means that combinations of several models with different degrees of freedom (DOF) are used to enhance the conformational sampling of biomolecules. The combination of all-atom (AA-) and coarse-grained (CG-) models is a typical example. In a "MultiScale Free Energy Landscape method" (MSFEL)[44-46] recently developed by one of us, a Go-like (C_α-atom based) model is first employed as the CG-model of proteins to perform an efficient conformational sampling using the smooth CG potential with low computational costs, which provides a rough but global Free Energy Landscape (FEL). Then, the obtained CG-FEL is refined through the AA-MD simulations with multiple umbrella samplings (USs),[47, 48] where several reference AA-structures are reconstructed based on C_α-atom structures sampled by the CG-MD simulation with structural databases for the backbones and rotamer libraries for the side chains. Finally, trajectories from the multiple USs are combined with the Weighted Histogram Analysis Method (WHAM)[49-51] to obtain the refined FEL. However, MSFEL has several weak points, such as a lack of contribution from hydrogen bonds, which are essential for forming secondary structure, in spite of its efficiency. In MSFEL, there is no coupling between CG- and AA-MD simulations, meaning that they are independently performed. In contrast, there are several hybrid methods that combine them through coupling interactions. One example of combined multiscale simulations is a MultiScale Essential Sampling (MSES).[52] In MSES, the conformational sampling in the AA-model is enhanced through a coupling with the accelerated essential dynamics described by the CC-model, where the AA- and CG-models are coupled by introducing an extra coupling potential. Here, a Hamiltonian Replica Exchange Method (HREM)[53, 54] can remove the biasing potential from the coupling term by exchanging the coupling parameters. In this strategy, there are no reconstructions from CG to AA structures and we do not take care of modelling both backbones and side chains. That is why MSES might be applicable to relatively large systems.

From this perspective, we introduce four enhanced conformational sampling methods for biomolecules, Parallel CaScade MD (PaCS-MD),[55] Fluctuation Flooding Method (FFM),[56] Outlier FLOODing (OFLOOD)[57] method, and TaBoo SeArch (TBSA)[58] method. Based on the above reviews of previously developed conformational sampling methods, we have designed simple, yet powerful, conformational sampling methods that are quite easy to implement and widely applicable to biomolecules based on distributed computing. A basic concept of our methods is as follows: (i) selection of initial seeds (initial structures) for MD simulations, (ii) conformational resampling of the selected seeds through restarting of short-time MD simulations initiated from them after initializing velocities, (iii) estimation of FEL by US with WHAM for reactive trajectories taken from a repetition of the cycle of (i) and (ii). The key points in our methods are how to select the candidates with high potential to make transitions as dominant seeds and to increase transition probabilities by reassignment of the initial velocities in restarting MD simulations. To extract the essential seeds from MD trajectories of (ii), appropriate "measures" are specified and the seeds are selected based on the measure. The four methods stated above utilize suitable measures to induce conformational transitions for given problems. The basic concepts of the four methods and detailed explanations of each method are given in Sec. 2. Practical applications of our methods are reviewed, including large domain motions upon the conformational transitions of proteins, conformational transitions of biomolecules upon ligand binding, and folding processes of proteins in Sec. 3. Finally, we briefly summarize the results in Sec. 4.

2. Methodology

2.1 Basic concept for extracting conformational transitions as biologically relevant rare events

As mentioned in the introduction, our methods are based on conformational resampling from reasonably selected "seeds" (initial structures for short-time MD simulations) for inducing conformational transitions. In our methods, the seeds for the conformational resampling are first selected by considering several appropriate "measures" that characterize the conformational transitions of biomolecules. Then, short-time MD simulations for the selected seeds at a given temperature are independently performed after initializing velocities. Because of the regeneration of the initial velocities, some seeds might happen to obtain sufficient kinetic energy to overcome the energy barriers among local energy minima. Actually, several past studies have reported the enhancement of the conformational sampling of proteins because of the regeneration of initial velocities in restarting MD simulations.[59-61] Figure 1 shows the flowchart of our methods through conformational resampling from the seeds. It shows a cycle of (i) selection of seeds based on the proper measures and (ii) conformational resampling from the selected seeds through the short-time MD simulations that might promote the conformational transitions of proteins. (iii) Then, FELs are evaluated along reactive trajectories obtained from a cycle of (i) and (ii) by adopting multiple US with the WHAM.

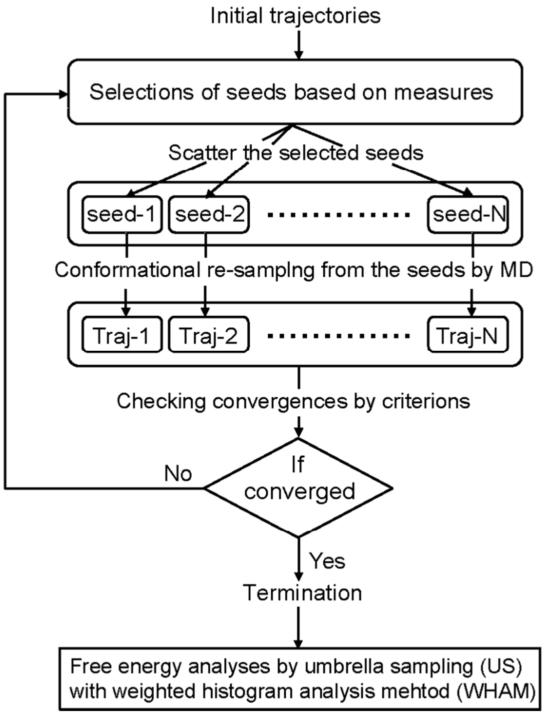
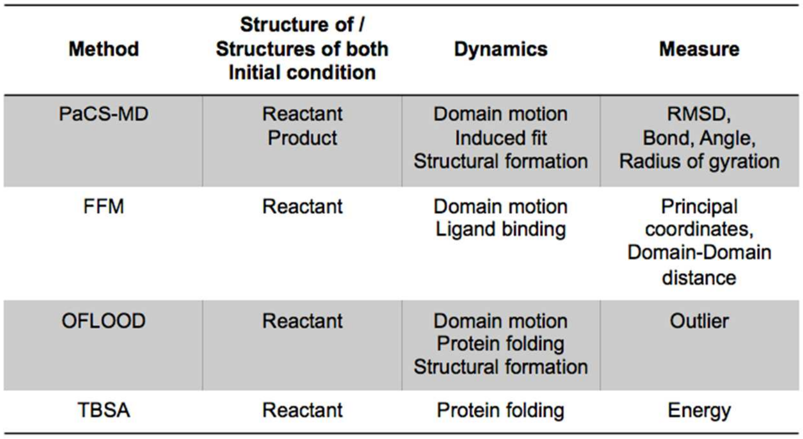

Initial trajectories
Selections of seeds based on measures
Scatter the selected seeds
seed-1
seed-2
seed-N
Conformational re-samplng from the seeds by MD
Traj-1
Traj-2
Traj-N
Checking convergences by criterions
If converged
No
Yes
Termination
Free energy analyses by umbrella sampling (US) with weighted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

Figure 1. Flowchart for selecting seeds and their conformational resampling. In the four methods (PaCS-MD, FFM, OFLOOD, and TBSA), the basic concept for enhancing the conformational transitions of biomolecules is common and the difference arises at the process of selection.

In these simple procedures, the crucial point is how to select seeds for the conformational resampling to extract effectively biologically relevant rare events, which also means how to define appropriately the measures characterizing the conformational transitions of interest. In our methods, definitions of measures are different in each method depending on the purposes of the research, meaning what kinds of biologically relevant rare events should be extracted under initial given conditions. To categorize measures, let us simply consider two cases. The first case is to extract biologically relevant rare events as the conformational transition pathways connecting two end-point structures, *i.e.*, a set of a reactant and a product is given *a priori*. The second case is to sample the conformational transition pathways or find the local and global energy minima of the biomolecules starting solely from a given initial structure. Of course, the latter case is a more difficult problem than the former. Herein, PaCS-MD is categorized in the former and FFM, OFLOOD, and TBSA in the latter. In Table 1, their features are given in view of initial conditions, their dynamics, and the measure. As seen in this table, the measures strongly depend on the purpose. In the following, we explain in detail how to define the measures and to extract the conformational transitions of interest for each method.



Method	Structure of / Structures of both Initial condition	Dynamics	Measure
PaCS-MD	Reactant Product	Domain motion Induced fit Structural formation	RMSD, Bond, Angle, Radius of gyration
FFM	Reactant	Domain motion Ligand binding	Principal coordinates, Domain-Domain distance
OFLOOD	Reactant	Domain motion Protein folding Structural formation	Outlier
TBSA	Reactant	Protein folding	Energy

Table 1. Features of each method depending on initial conditions, dynamics, and measures.

2.2 Parallel CaScade Molecular Dynamics (PaCS-MD)

Parallel CaScade MD (PaCS-MD) is a conformational path generation method under a condition that a set of the reactant and the product is given *a priori*. In PaCS-MD, for instance, the seeds are selected based on a structural similarity with respect to the product and followed by a conformational resampling from the selected seeds, so that the selected seeds move gradually closer to the product. For the flowchart of PaCS-MD, see Fig. 1 in reference.[55] Repeating a cycle of selection of the seeds and their conformational resampling through short-time MD simulations (several hundred picoseconds) at a given temperature, PaCS-MD might generate candidates of the conformational transition pathways connecting the reactant to the product. As an example of the measure based on the structural similarity to the product, root mean square deviation (RMSD) from the product (the target structure) is the most general choice. After several short-time MD simulations starting from several seeds, RMSDs are evaluated and ranked. By this specification, snapshots with small RMSDs among all trajectories might be selected as reasonable seeds for the next short-time MDs. Other measures might also be considered depending on the purpose of the simulations. The cycles of selection and short-time MD simulations are subsequently repeated until highly ranked snapshots sufficiently reach the product, say RMSD < cut-off, where cut-off is a threshold for the termination. Finally, multiple trajectories generated from the cycles are joined as "reactive trajectories" connecting the reactant to the product, which corresponds to forming "quasi-conformational transition pathways". The reactive trajectories are obtained by tracing the surviving seeds from the last to the initial cycles. These quasi-conformational transition pathways generated by PaCS-MD might be relatively close to, but slightly different from, the real pathway. Therefore, it might be better that the reactive trajectories are refined through appropriate path-sampling methods such as TPS or String Method. As initial guesses for these path-sampling methods, the reactive trajectories are refined to optimal conformational transition pathways. Alternatively, as mentioned in reference,[55] the realistic pathways should be quantitatively evaluated by FEL analyses through USs using multiple reference structures along the quasi-conformational transition pathways, as described in Sec. 2.6.

2.3 Fluctuation Flooding Method (FFM)

The Fluctuation Flooding Method (FFM) is a general extension of PaCS-MD. In contrast to PaCS-MD, FFM requires only a single structure as an initial reactant. FFM uses degrees of anisotropic fluctuations found in biomolecules, which are defined through several analytical methods, such as Principal Component Analysis (PCA),[62, 63] to enhance the conformational transitions under the hypothesis that highly fluctuating structures along anisotropic modes tend to undergo conformational transitions with high probabilities. In general, PCA is performed by diagonalizing a variance–covariance matrix $\boldsymbol{C}$, which is taken from relatively long-time MD simulations for an ensemble average, and defined as

$$\boldsymbol{C} = \left\langle \left(\mathbf{q} - \langle \mathbf{q} \rangle\right)^T \left(\mathbf{q} - \langle \mathbf{q} \rangle\right) \right\rangle, \tag{1}$$

where $\mathbf{q}$ is a mass-weighted coordinate vector of a given protein consisting of N atoms, and the brackets mean the ensemble average. PCA gives a set of eigenvectors $\mathbf{u}_i$ and eigenvalues λ_i (i = 1, 2, … , 3N) as the i-th principal mode (PM) and as their contributions to each PM, respectively. Herein, PMs are used as a general concept for describing "all fluctuation modes" that are obtained as eigenvectors by diagonalizing a variance–covariance matrix. In contrast, principal components are used as "specific collective modes" that are taken from PMs as a reference frame relevant to biological functions. For example, in domain motions of proteins, the top m PMs with large λ_m values, *i.e.*, PM_1, PM_2, … , PM_m, correspond to large-amplitude collective motions reflecting the anisotropic fluctuations of the proteins. This means that the principal coordinates (PCs) projected onto the top m PMs, PC_1, PC_2, … , PC_m, are the collective coordinates for describing the anisotropic fluctuations relevant to biological functions. In FFM, highly fluctuating structures along anisotropic PMs are regarded as the seeds that have high potential to transit in the conformational resampling. The seeds are selected by ordering magnitudes of the anisotropic fluctuations defined by the absolute values of PCs after the projections of MD trajectories onto anisotropic PMs. Then, the conformational resampling of highly fluctuating structures is repeated via short-time MD simulations. FFM consists of a cycle of the following two schemes: (i) selection of seeds with the minimum and maximum PCs projected onto the top m PMs and (ii) conformational resampling from the selected seeds via regeneration of initial velocities in restarting MD simulations. These cycles are repeated until the conformational transitions have converged. For the flowchart of FFM, see Fig. 1(b) in reference.[56] For the termination of FFM, convergence of distributions projected onto each PM should be checked between adjacent cycles, *i.e.*, the k – 1-th and the k-th cycles. To check whether each distribution converges or not, the following index, σ_k, might be helpful,

$$\sigma_k = \frac{\int d\xi[\rho_k(\xi) - \rho_{k-1}(\xi)]^2}{\int d\xi[\rho_{k-1}(\xi)]^2}, \tag{2}$$

where $\rho_{k-1}(\xi)$ and $\rho_k(\xi)$ are cumulative distributions projected onto the RC, ξ, at the k – 1-th and the k-th cycles. FFM will be terminated if the index defined by Eq. (2) is sufficiently small, *i.e.*, $\sigma_k < \varepsilon$, where ε is a threshold to judge the termination.

2.4 Outlier FLOODing (OFLOOD) method

The Outlier FLOODing (OFLOOD) method is an enhanced conformational sampling method that uses outliers, which are defined as snapshots of sparse distributions on the conformational space spanned by a set of RCs and detected by clustering methods from the obtained trajectories. When investigating biological functions, appropriate RCs should first be defined to characterize biological processes. Then, states of biomolecules are projected onto the conformational space spanned by RCs, generating high-dimensional distributions on the subspace. Here, the key point is how to treat the high dimensionality of the conformational distributions according to the increasing DOF of proteins. To extract biologically relevant rare events from the high-dimensional distributions effectively, clustering techniques automatically detect major and minor components as the coarse-grained information, so that it is suitable for directly and systematically treating high-dimensional distributions.

Herein, stable states of a given protein are characterized as dense distributions, *i.e.*, *clusters* extracted by clustering analyses. In contrast, the *outliers*, which belong to complementary components to the clusters, are found in sparse distributions located typically at the edge of the clusters. When two or more clusters are close, some outliers at an overlapped region of plural clusters might be detected near transition states. If the outliers are extensively selected as the seeds, some seeds cross over the barrier to find a new local minimum by chance and others expand regions of the sparse distributions, as demonstrated later. That is why the outliers are regarded as appropriate seeds that have high potential to transit from one cluster to another in the conformational resampling. Because of conformational resampling from the outliers, biologically relevant rare events might be stochastically induced through jumping among the clusters through the regeneration of initial velocities in restarting short-time MD simulations. In OFLOOD, a cycle of the following schemes is repeated: (i) extraction of outliers as the seeds by clustering of MD trajectories and (ii) conformational resampling from the selected outliers by restarting MD simulations. In every cycle, the clustering is updated using all trajectories from the past cycles to extract new outliers, which are used for the next seeds. The conformational resampling is continued until the conformational search converges sufficiently. For the termination of OFLOOD, convergence of cumulative distributions projected onto RCs should be checked between adjacent cycles based on Eq. (2) until σ_k becomes sufficiently small.

2.5 TaBoo SeArch (TBSA) method

The TaBoo SeArch (TBSA) method is an enhanced conformational sampling method based on information of distributions projected onto RCs. In general, states of a system are defined by frequent peaks in distributions, *i.e.*, histograms of conformational spaces projected onto RCs. To obtain the histogram on conformational space, the energy of a given system is one of the general RCs. Once the histogram along the energy space is obtained, the states of the system with low frequencies are judged from the magnitude of the histogram. In the calculation of the histogram, the conformational sampling of these sparse regions, which correspond to "tails" of the histogram, is necessary to obtain a correct statistical ensemble. However, limited time-scale MD simulations fail to sample sufficiently these sparse regions. To remedy the insufficiency in the conformational sampling, TBSA intensively performs conformational resampling of the states with low frequencies using a series of short-time MD simulations starting from reasonably selected seeds. In TBSA, the seeds are randomly selected using an inverse histogram as their probability weight derived from the original histogram. This means that TBSA forces the states with low frequencies to be intensively resampled, improving the insufficiency of the conformational sampling at the tails of the histogram. Based on the above strategy, a cycle of the following schemes is repeated: (i) selection of seeds for the conformational resampling of sparse regions based on its inverse histogram $\rho_{\mathrm{inv}}(E)$ as a function of energy E,

$$\rho_{\mathrm{inv}}(E)=\alpha(E)\left[\rho_{\max}-\rho(E)\right]\theta(E_{\max}-E)\theta(E-E_{\min}), \tag{3}$$

where $\rho(E)$, $\rho_{\max}$, and $\alpha(E)$ are the original histogram, the peak with the maximum value, and the normalization factor for adjusting the total number of the seeds, respectively. The step functions, θ, in the above equation ensure that the inverse histogram at unvisited regions is defined to be zero, so that selection of the seeds is random with the weight being proportional to the normalized inverse weight within a domain of the minimum and maximum energies, $\left[E_{\min}, E_{\max}\right]$, which are updated at every cycle. (ii) The conformational resampling by restarting short-time MD simulations is initiated from the selected seeds through regeneration of initial velocities and followed by updating the inverse histogram for the next selection. For the schematic concept of TBSA, see Fig. 1(b) in reference.[58] To judge whether TBSA is terminated or not, convergence of the histogram projected onto the energy space should be checked. As a convergence criterion in actual applications, the averaged error between the k- and k + 1-th distribution $\left|\rho_{k+1}(E)-\rho_k(E)\right|/N_{\mathrm{bin}}$ is imposed, where N_{bin} is the number of bins used to describe the histogram.

2.6 FEL analyses by US with WHAM

Once the conformational sampling by each method is performed, quasi-conformational transition pathways are obtained as reactive trajectories, which correspond to coarse-grained transitional pathways, as mentioned in Sec. 2.2. Strictly speaking, they might be different from the true pathways. Therefore, for their refinement, these reactive trajectories should be quantitatively evaluated through FEL analyses, allowing us to obtain a better description of the transition pathways judging from free energy values. For the FEL analyses, a combination of the USs with the WHAM might be an efficient strategy. In general, if R reference structures for the multiple US are selected along the reactive trajectories, the conformational sampling for each reference will be intensively performed using the following umbrella potentials as a function of coordinates, $\mathbf{r}$,

$$V_s^{\mathrm{bias}}(\mathbf{r})=\frac{k_s}{2}\left(\mathbf{r}-\mathbf{r}_s^{\mathrm{ref}}\right)^2 \quad (s=1,2,\cdots,R), \tag{4}$$

where $\mathbf{r}_s^{\mathrm{ref}}$ and k_s are coordinates of the s-th reference and the spring constant for the harmonic restraint. After that, biased probability densities obtained from the multiple USs, $\rho_s^{\mathrm{biased}}(\mathbf{r})$, are unbiased to obtain unbiased probability densities $\rho_s^{\mathrm{unbiased}}(\mathbf{r})$,

$$\rho_s^{\text{unbiased}}(\mathbf{r}) = \frac{e^{\beta V_s^{\text{bias}}(\mathbf{r})} \rho_s^{\text{biased}}(\mathbf{r})}{\left\langle e^{\beta V_s^{\text{bias}}(\mathbf{r})} \right\rangle_{\text{biased}}}, \tag{5}$$

where $\beta = 1/k_{\mathrm{B}}T$ is defined as the inverse temperature defined in terms of the Boltzmann constant, k_{B}, and the absolute temperature, *T*. The unbiased probability densities are joined as an optimal probability density with the WHAM as follows:

$$\rho(\mathrm{r}) = \sum_{s=1}^{R} w_s(\mathrm{r}) \rho(\mathrm{r})_s^{\text{unbiased}}(\mathrm{r}), \tag{6}$$

where weights $\{w_s(\mathbf{r})\}$ $(s = 1, 2, \cdots, R)$ for each unbiased probability density are determined so that statistical errors σ^2 should be minimized under a normalization condition for the weights as,

$$\begin{cases} \dfrac{\partial\left(\sigma^2\left[\rho(\mathbf{r})\right]\right)}{\partial w_s(\mathbf{r})} = 0 \\ \displaystyle\sum_{s=1}^{R} w_s(\mathbf{r}) = 1 \end{cases} . \tag{7}$$

Finally, FEL, $F(\xi)$, measured from the origin, $F(\xi_0)$, is defined as the negative logarithm of the probability density projected onto an RC, ξ,

$$F(\xi) - F(\xi_0) = -\frac{1}{\beta} \ln \rho(\xi) . \tag{8}$$

3. Results and Discussion

In this section, several applications using our methods are reviewed. Herein, our methods were applied to biologically important phenomena depending on the purposes of the research, such as extraction of domain motions of proteins, protein folding, and conformational transitions upon ligand binding.

3.1 Domain motion analyses by PaCS-MD and FFM

As a demonstration of extraction of the conformational transitions via domain motions of proteins, let us consider bacteriophage T4 lysozyme (T4L) (164 residues) explicitly in a water environment. For the dynamics of T4L, several preceding studies[64-66] have proved that the collective motions in the hinge angle between two domains are induced as open–closed conformational transitions. The structures of both open (PDBid: 150L[67] as the mutation, M6I) and closed (PDBid: 2LZM[68] as the wild type) forms have been determined using X-ray crystallography. Because a set of the reactant (open form) and the product (closed form) is known *a priori*, PaCS-MD could be applied to the conformational transitions of T4L through the generation of the conformational transition pathways between these end-point structures. As a measure to select appropriate seeds for the conformational resampling, RMSDs from the target structures were considered. Based on RMSDs from the targets, snapshots with the 10 lowest RMSDs were selected from trajectories of short-time MD simulations in each cycle. By repeating the cycles of selection and conformational resampling, they were gradually brought close to the targets. For the profiles of RMSDs of the seeds in each cycle, see Fig. 3(a)–(e) in reference.[55]

In this review, we demonstrate another approach for PaCS-MD to generate the conformational transition pathways between the open and closed forms of T4L. In PaCS-MD simulations for the conformational transition from the open to the closed state, the product structure for the open form T4L was modelled from a mutant X-ray structure (PDBid: 150L), where a mutated back structure from 150L to the WT was constructed, because the X-ray structure of the open form WT has not been determined by experiment. In actual simulations, the constructed open structure was equilibrated through a short-time MD simulation. Finally, the equilibrated snapshot was employed as a model structure for the PaCS-MD from the closed to the open transition. To neutralize the system, eight chloride ions were also added, amounting to a total of 29252 atoms. For the closed form (PDBid: 2LZM), almost the same recipe as for the open form was adopted to yield the model system. All MD simulations were performed using the PMEMD module of AMBER 11[69] with the AMBER 99SB force field.[70] In general, the collective motions of proteins such as domain motions are described by a small number of PMs, because the contributions of a few PMs in higher modes are dominant because of the isotropic fluctuations of the proteins. As a measure to select the seeds for the conformational resampling, PCs projected onto PMs were chosen instead of RMSD, where PMs were determined as eigenvectors $\mathbf{u}_i (i = 1, 2, \ldots , 3N)$ obtained by diagonalizing the variance–covariance matrix, $\boldsymbol{C}$, defined in Eq. (1). To determine the PMs, conventional 10-ns CMD simulations using the *NVT* ensemble (T = 300 K using the Berendsen thermostat)[71] were started from the equilibrated open and closed forms. Then, each trajectory was joined (total 20 ns) and an averaged structure was calculated as a reference, which corresponds to $\langle \mathbf{q} \rangle$ in Eq. (1). In this determination of PMs, the average structure is assumed as a transition state between the open form and the closed form. To check how many PMs are required to describe the collective motions of T4L, we calculated an index, κ, defined by the accumulated eigenvalues until the *i*-th PM as

$$\kappa = \sum_{i=1}^{n} \lambda_i \Big/ \sum_{m=1}^{3N} \lambda_m . \tag{9}$$

In this demonstration, the index coming from the top 20 PMs covered 90% of the overall eigenvalues. This evidence means that the top 20 PMs might be enough to describe the collective motions upon the open–closed conformational transition of T4L. To define the measure in PaCS-MD, the Inner Product (*IP*) within the 20 dimensions $\left(N_{\text{dim}} = 20\right)$ was considered to describe the correlation of PMs between snapshots of MD trajectories and the products. Here, the *IP* value is defined as:

$$IP = \sum_{i=1}^{N_{\text{dim}}} \overrightarrow{PC}_i^{\text{snapshot}} \cdot \overrightarrow{PC}_i^{\text{product}} , \tag{10}$$

where $\overrightarrow{PC}_i^{\text{snapshot}}$ and $\overrightarrow{PC}_i^{\text{product}}$ are the *i*-th PC of a snapshot of the MD trajectories and the product, respectively. Snapshots with high *IP* values are intensively selected as the seeds for conformational resampling. PaCS-MD is terminated when *IP* values in the 20-dimensional space become sufficiently close to one.

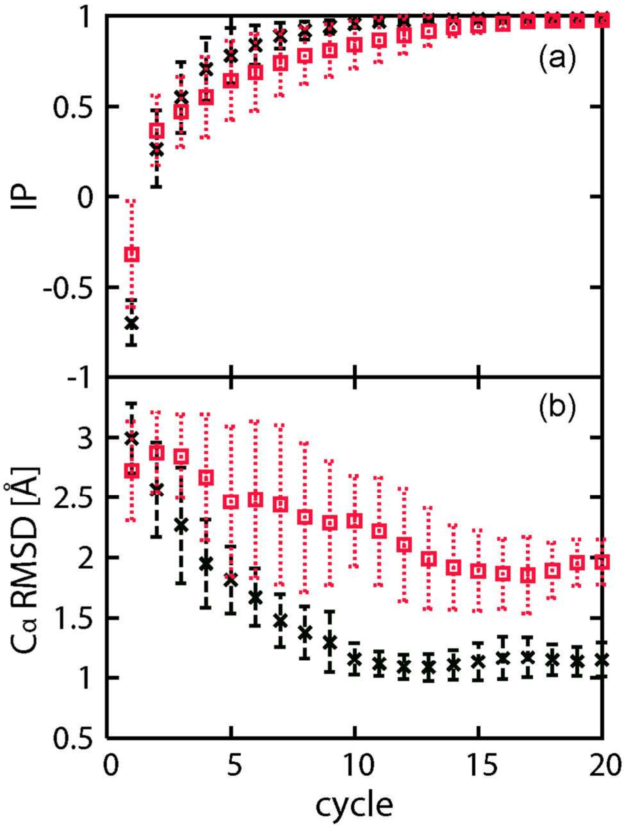


Figure 2. Profiles of averages with error bars according to the cycles, for (a) *IP* value and (b) C_α RMSD with respect to each product. The black crosses and red squares correspond to the averages upon the conformational transitions from the open to closed forms and vice versa, respectively.

For both directions (the conformational transition from the open to the closed forms, and vice versa), 10 distinct PaCS-MD trials were performed. In each trial, PaCS-MD was continued for 20 cycles. In each cycle, seeds with the 10 top *IP* values were selected as initial structures for the short-time (100-ps) MD simulations. For the conformational transition from the open to the closed forms, the relaxed closed form in an explicit water environment was regarded as the product. On the other hand, for the inverse conformational transition, the mutated-back relaxed open form in an explicit water environment was regarded as the product. Figure 2(a) shows the profiles of *IP* values averaged over the 10 distinct trials with their standard deviations (until the 20th cycle). In each direction, *IP* values between the selected seeds and the products converged sufficiently to one during the 20 cycles. As shown in Fig. 2(b), the corresponding RMSDs measured from the products also reasonably converged to low values in each direction with increases in the *IP* values. Herein, one has to note that RMSD of the conformational transition from the closed to the open state converged to relatively large values (≈2.0 Å) compared with the inverse transition because the mutated-back open form was selected as the model structure to measure RMSD from the open state instead of the WT open form. This evidence indicates that 20 cycles are enough for the iterations to converge. To obtain greater knowledge on the transition pathway, the reactive trajectories were projected onto the two-dimensional subspace spanned by PC1 and PC2, as depicted in Fig. 3. According to these figures, there exist two characteristic transition pathways spanned on the subspace. These pathways might be used as initial guesses to the well-established transition path sampling methods reviewed in the introduction. In addition, these coarse-grained conformational transition pathways generated by PaCS-MD were quantitatively evaluated through the FEL analyses using multiple USs along RCs, as shown in our previous study (see Fig. 9(c) in reference[55]). According to our previous study,[55] the conformational transition pathway presumed by Fig. 3(a) might be close to a realistic route (see Fig. 9(c) in reference[55]). However, another conformational transition pathway of Fig. 3(b) seems not to be far from that of Fig. 3(a). Therefore, the multiple US using reactive trajectories constructed from Fig. 3(b) will converge to Fig. 3(a) through the FEL calculation because they have a higher free energy profile than the realistic pathway judging from Fig. 9(c) of reference.[55]

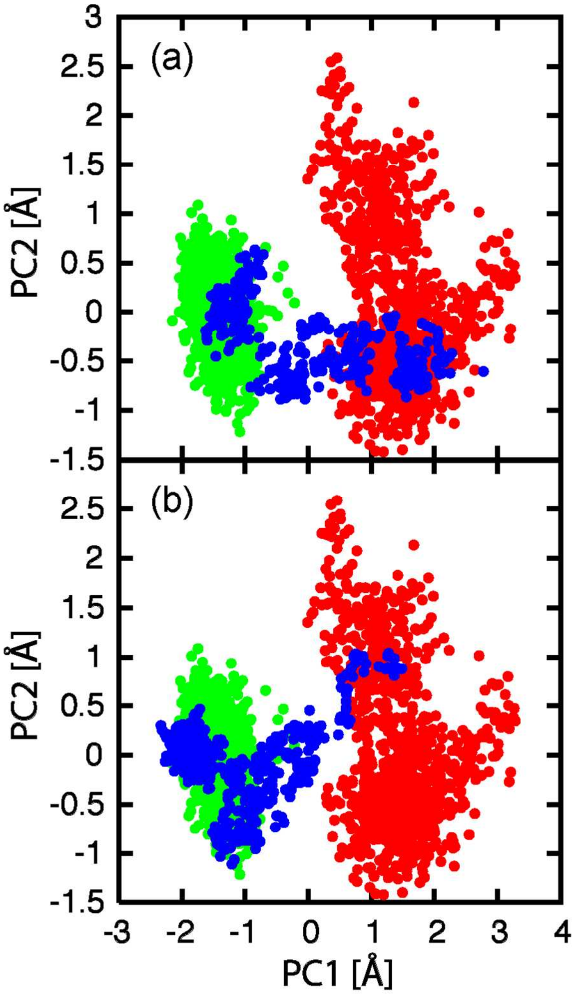


Figure 3. Projections of the reactive trajectories onto the subspace spanned by PC1 and PC2 according to the cycles for the conformational transition (a) from the open to closed forms and (b) vice versa (blue points). The red and green points correspond to the projections of trajectories obtained from the 10-ns CMD simulations using the *NVT* ensemble (*T* = 300 K) started from the relaxed open and closed forms, respectively. The projections of the target structures (products) in PaCS-MD are as follows: mutated back to WT and relaxed open form: (0.7, 1.3), WT closed form: (–1.5, 0.3).

As an example of the FEL calculation combined with PaCS-MD, an application to Glutamine Binding Protein (GlnBP) is presented here. GlnBP is one of the periplasmic proteins, consisting of 226 residues and amounting to 3535 atoms. When the ligand (glutamine) binds to a cleft between two domains in the open form, large hinge-bending movements are induced towards the closed form. The structures of the open-apo (PDBid: 1GGG)[72] and the closed-holo (PDBid: 1WDN)[73] forms have been determined using X-ray crystallography.

Note here that the conformational transitions of the apo-type (without the ligand) were investigated. To model initial structures, structures of the open-apo and closed-apo forms were prepared from these X-ray structures, where for the latter the ligand was removed from the closed-holo form, because the X-ray structure of the closed-apo form is not available. An explicit water environment was modelled with the TIP3P water, and one chloride ion was added to neutralize the systems, finally amounting to 33533 atoms as solvated model systems. After energy minimization, short-time (300-ps) MD simulations were started from the model structures for adjusting volumes of the systems using the *NPT* ensemble (P = 1.0 atm, T = 300 K using the Berendsen thermostat). Finally, the last snapshots for both systems were regarded as the products. After the volume adjustment with a short-time simulation using the *NPT* ensemble, the environment was switched to *NVT* (T = 300 K). As the measure in PaCS-MD, an *IP* value between snapshots and the products was used. To determine PMs by PCA, relatively long-time (10-ns) MD simulations were independently started from the relaxed open-apo and closed-apo forms using the *NVT* ensemble (T = 300 K), and then joined (20-ns) trajectories were used to define the variance–covariance matrix, as was done for T4L above. In each cycle of PaCS-MD, the 10 top seeds with high *IP* values were selected. Then, short-time (100-ps) MD simulations were independently restarted from the selected snapshots. Finally, the cycle was repeated for 30 cycles (total 30 ns computational time, 10 seeds × 100-ps MD simulations × 30 cycles).

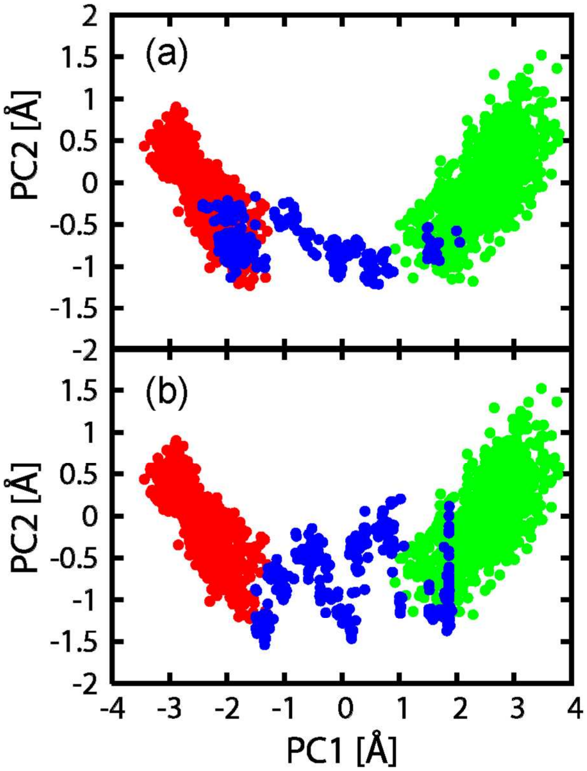


Figure 4. Projections of the reactive trajectories onto the subspace spanned by PC1 and PC2 according to the cycles (the blue points). (a) The directions from the open form to the closed form and (b) vice versa. The green and red points represent the projections of trajectories obtained from the 10-ns CMD simulations using the *NVT* ensemble (T = 300 K) for the open and closed forms, respectively.

In this demonstration, PaCS-MD was performed for 30 cycles by referring the *IP* value between the seeds and the products for both directions of the conformational transitions (the open form to the closed form and vice versa). *IP* values sufficiently converged to one during the 30 cycles. Figure 4 shows the projections of the selected seeds during the 30 cycles for both conformational transitions onto the subspace spanned by PC1 and PC2 (Fig. 4(a) for the open to the closed and Fig. 4(b) for the closed to the open). To evaluate quantitatively the generated reactive trajectories, FEL calculations by multiple USs combined with WHAM were performed, where references for the multiple USs were selected along the reactive trajectories and restraints on C_α co-ordinates of the multiple references (force constant = 0.0001 kcal/(mol Å^2)) were added in the multiple USs. In total, 100 references were selected from the reactive trajectories, so that they would be uniformly distributed along the transition pathways. The FEL of GlnBP projected onto the subspace spanned by PC1 and PC2 is shown in Fig. 5. Judging from the FEL calculation, several metastable states during the conformational transitions were observed as local free energy minima along the transition pathways (closed, semi-closed, transition, and open in Fig. 5). The semi-closed form sampled by this method might have an important role in the ligand binding process in the holo-state whether these processes occur via the induced-fit or the populated shift. Actually, these types of intermediate states have been reported in preceding studies, such as Lysine-, Arginine-, Ornithine-binding protein (LAO) and Maltose-Binding Protein (MBP).[74, 75] In addition to these apo-type simulations, holo-type simulations with the ligand (glutamine) might enable us to understand the allosteric conformational transitions upon ligand binding. These simulations will be presented elsewhere.

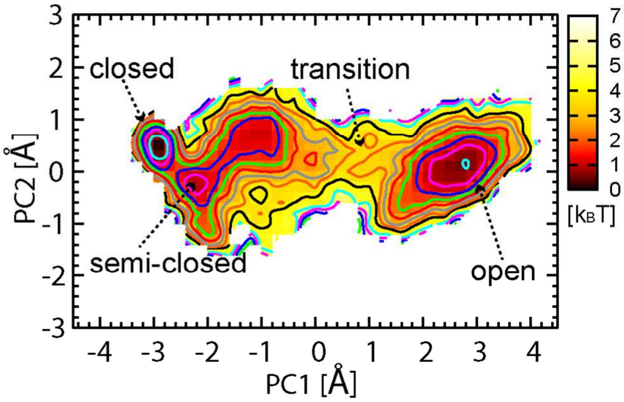


Figure 5. FEL of (apo) GlnBP projected onto the subspace spanned by PC1 and PC2 through multiple USs and the WHAM. The representative states are indicated with dashed arrows as closed, semi-closed, transition, and open states. The contour lines are drawn every 0.5 k_BT.

In contrast to PaCS-MD, which needs structures of both a product and a reactant *a priori*, FFM only needs that of the reactant to induce automatically conformational transitions towards other metastable states. Because one often encounters problems where only one structure is available from X-ray crystallography, the FFM has an advantage over PaCS-MD for application to these systems. To emphasize the difference between FFM and PaCS-MD, we here demonstrate the FFM simulations of T4L by referring to the open form only. For extraction of the conformational transition starting with the open structure of T4L, the relaxed open form (150L) used in the previous paragraph was considered as a reactant. First, a canonical MD simulation was performed from the relaxed open structure for 10 ns using the *NVT* ensemble (T = 300 K using the Berendsen thermostat) to determine PMs for the open forms. Note that these PMs are different from those adopted in the previous paragraph. Here, the accumulated contribution coming from PM1 and PM2 totalled 70.8%, which might be sufficient to describe the collective motions of T4L. Therefore, PM1 and PM2 were considered to select seeds that have large amplitudes as anisotropic fluctuations. Based on the projections of MD trajectories onto these PMs, the snapshots that have minimum and maximum PCs were selected as seeds for the conformational resampling. In the actual selection of the seeds, the PM to be enhanced was randomly selected from PM1 or

PM2. In each cycle, a total of 10 seeds was selected and followed by the conformational resampling from them through short-time (100-ps) MD simulations via regeneration of initial velocities. As a reference, CMD simulations using the *NVT* ensemble (*T* = 300 K using the Berendsen thermostat) were also performed from the relaxed closed forms for 10 ns. In each cycle of FFM, highly fluctuating snapshots were detected as five sets of seeds (snapshots) with the minimum and the maximum PCs projected onto a randomly selected PM from PM1 and PM2, *i.e.*, a total of 10 seeds. Then, these 10 snapshots were resampled through restarting the short-time (100-ps) MD simulations in each cycle. The cycle was repeated to 100 cycles for a strict check of the convergence, amounting to a total of 100 ns computational time (10 seeds × 100-ps MD simulation × 100 cycles).

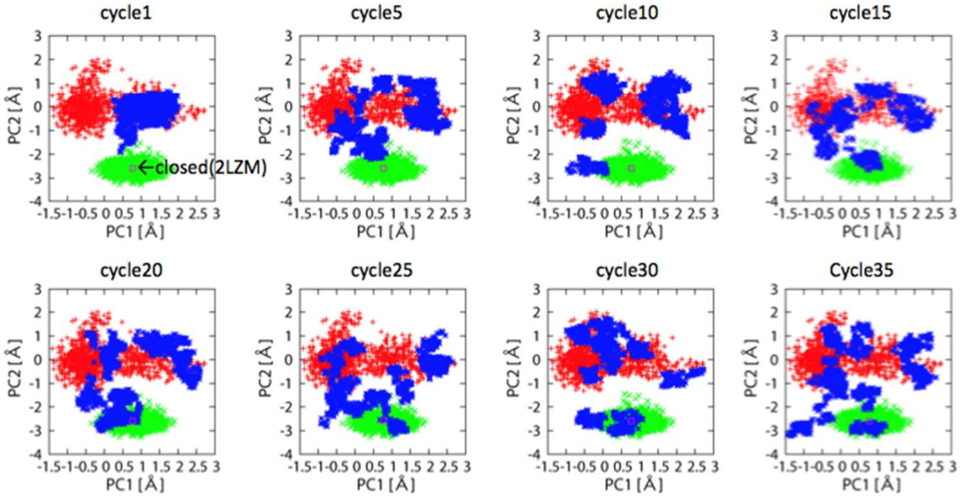


Figure 6. Projections of the snapshots obtained from FFM onto the subspace spanned by PC1 and PC2 according to the cycles as blue points. The red and green points correspond to the projections obtained from the 10-ns CMD simulations using the *NVT* ensemble (*T* = 300 K) started from the relaxed open and closed structures. The square (magenta) indicates the projection of the X-ray structure of the closed state.*

The blue point in Fig. 6 shows the projections of snapshots obtained from FFM onto the subspace spanned by PM1 and PM2 every five cycles. On the other hand, the red and green points correspond to those obtained from 10-ns CMD simulations for the open and closed forms, respectively. As is clearly shown in Fig. 6, FFM enhanced the conformational transitions from the open to the closed forms within the 15th cycle (total 15-ns computational time) because the projections (blue points) of snapshots obtained from FFM sufficiently covered the projections (green points) of the trajectory obtained from the CMD simulation started from the relaxed closed form including the X-ray structure indicated by the square in Fig. 6. For the structural similarity, the superposition of the structures sampled by FFM (red) and determined by the X-ray analysis (black) is shown in Fig. 3(b) in reference[56] (the minimum RMSD was 0.89 Å). Furthermore, according to the cycles, FFM broadly sampled not only the transitional region around the open and closed forms in the subspace spanned by PM1 and PM2, but also the marginal regions with high free energies, such as edges of the FEL. In our previous study, the FEL was quantitatively evaluated using the multiple USs combined with WHAM for reactive trajectories of T4L generated by FFM to yield reliable conformational transition pathways connecting the open and the closed forms. For details of the FEL of T4L, see Fig. 4(a) in reference.[56] For the FEL analysis, 100 multiple USs were performed for 1 ns under harmonic restraints with respect to C_α atoms of the references defined by Eq. (4),

*Figure 6 has already been published in reference.[56] However, the layout is different from the original layout.

amounting to a total of 100 ns computational time (100 references × 1 ns). As a comparison with the CMD simulations, a brute-force 1-μs CMD simulation was performed from the relaxed open form. Even if the μs-order CMD simulation was employed, it is quite difficult to extract the conformational transition from the open to closed states (see Fig. 4(b) in the reference).[56] This evidence clearly indicates that the conformational sampling efficiency of the FFM is remarkably higher than that of the brute-force CMD.

As a comparison of the conformational sampling efficiency of FFM, we compared our results with a recently proposed enhanced conformational sampling method, referred to as Accelerates Collective Motions of a protein obtained from PCA (ACM-PCA).[76] In ACM-PCA, several hundred picosecond-order MD simulations are carried out for the target protein and subsequently PCA is performed on each trajectory, providing reliable directions of domain motions relevant for biological functions. Finally, motions along the collective modes are accelerated by coupling them to a thermostat with a high temperature. A cycle of the above procedures is repeated to enhance the conformational sampling. Herein, the concept of ACM-PCA might be similar to that of FFM, in both of which snapshots/modes with high potentials to transit are characterized by PCA and accelerated by controlling velocities. In FFM, essential snapshots are extracted by referring projections of trajectories to PCs determined from PCA, and then the conformational transitions are stochastically accelerated by regeneration of initial velocities in restarting the MD. In contrast, ACM-PCA employs the thermostat coupling to enhance the collective motions of proteins. Therefore, the way to accelerate snapshots relevant for transitions might be different for the two methods. As a demonstration of ACM-PCA, this method was applied to T4 lysozyme, which is the same situation in our demonstration. In ACM-PCA, starting from the WT closed form, a broad conformational sampling including the conformational transition to the open form was achieved within 20 ns. However, the inverse transition from the open to closed state was not obtained with ACM-PCA. The open to closed transitions of T4L seem quite difficult to extract. Actually, we performed 1-μs CMD simulation with explicit water using the *NVT* (T = 300 K) starting from the open form. In this trial simulation, we could not find the transition from the open to closed form (see Fig. 4(b) of the reference).[56] In contrast, FFM successfully extracted the rare event within 20 ns without using any information about the closed form; we used only PCs obtained from 10-ns MD simulations starting from the open state. The above discussion indicates the high conformational sampling efficiency of FFM, and this method has a quite high potential to extract biologically rare events.

3.2 Analyses of the induced-fit process after ligand binding by PaCS-MD

As one of the biologically relevant rare events, induced-fit processes after ligand binding have received attention from both experimental and theoretical points of view because these processes are often related to enzymatic activities. One of the nylon-oligomer hydrolases called NylB, which is found in the *Arthrobacter sp.* KI72,[77] degrades amide bonds in linear nylon-oligomers,[78] whose reaction mechanism was recently revealed by CMD and first-principles MD by us, and interfragment analyses were performed to investigate ligand binding processes.[79-81] After the ligand binding to a cleft, a specific residue Y170 in NylB, which belongs to a loop segment (Asn166–Val177), binds to the substrate, 6-aminohexanoate dimer (Ald), forming a hydrogen bond with it. During the induced-fit process, the loop region moves significantly towards the substrate. A site-directed mutagenesis study showed that the Y170F mutant, where Tyr is replaced by Phe, has less enzymatic activity than the wild type (WT) does. To investigate the difference between the WT and the Y170F mutant in the induced-fit processes, we employed PaCS-MD to calculate FELs of both the WT and the Y170F mutant. [82, 83]

Initially, we prepared the prototype structure for the WT enzyme–substrate complex (closed form) from the X-ray structure available at the Protein Data Bank (PDBid: 2ZMA).[84] In this structure, the Ser112 residue is replaced by Ala, so that we replaced Ala112 with the original Ser112 residue and added all of the missing amino acid residues to obtain a closed-form WT model. The open form was also obtained from the X-ray structure (PDBid: 2ZM0).[85] Analogously, the open and closed forms of the Y170F mutant were reconstructed by replacing Tyr170 with Phe170 in the corresponding WT models. The WT system here consists of the enzyme–substrate complex solvated in water (described by a TIP3P force field) with Na^+ counter ions to neutralize the system, amounting to 41403 atoms. The systems were then equilibrated by CMD for about 10 ns using an *NPT* ensemble (P = 1 atm and T = 300 K using the Berendsen thermostat). For the PaCS-MD simulations, the reactant and the product structures for both the WT and the Y170F mutant were extracted from the last snapshot of each 10 ns CMD simulation using the PMEMD module of the AMBER 12 program suite.[86] The force field adopted for the amino acid residues was a standard Amber ff99,[87] while the force field for Ald was generated with the AMBER GAFF by the Antechamber program.[79] PaCS-MD simulations were performed for 25 cycles starting from 10 different initial seeds sampled on the basis of the measure of a partial RMSD of the heavy atoms in the loop segment (Asn166–Val177). To evaluate the free energy change because of the induced fit, we adopted multiple US combined with the WHAM, where 2000 reference structures randomly extracted from PaCS-MD were used. For the multiple US, restraints with a force constant k = 0.5 kcal/(mol Å^2) were imposed on the heavy atoms in the loop segment. The FELs were then projected onto two specific distances: the separation between the C_{δ} atom of Glu168 and the C_{α} atom of Ser217, d_1, and the separation between the H_{ζ} atom of Phe170 (replacing the O_{ζ} atom of Tyr170 in the WT) and the N atom in the amide bond of Ald393, d_2, where the definitions of these distances are given in Fig. 7(a).

To elucidate the mutational effects of Y170F on the induced-fit process, FELs for the WT and the Y170F mutant are depicted in Figs. 7(b) and 7(c). For each FEL, the left bottom region corresponds to the closed form (the global minimum) and the upper right region to the open form. According to these figures, two major features are found. One is that the number of (meta-) stable structures is different, *i.e.*, six for the WT and four for the Y170F mutant. The other is the difference in the slope near each local minimum. In the case of the Y170F mutant, the areas of the FEL around the minima at (d_1, d_2) = (10.0 Å, 14.0 Å) and (d_1, d_2) = (9.0 Å, 7.5 Å) in Fig. 7(c), but also the other two (d_1, d_2) = (7.1 Å, 5.5 Å) and (d_1, d_2) = (4.8 Å, 5.2 Å), are rather wide and characterized by not very steep slopes with respect to the WT. The final minimum, (d_1, d_2) = (4.9 Å, 5.3 Å), is a bit far from that obtained for WT, reflecting a difference in the O_{ζ} atom of Tyr170 and the H_{ζ} atom of Phe170. Despite this important difference in the shape of the FESs, the highest free energy barrier in the FEL of the Y170F mutant (2.4 kcal/mol) is nearly identical to that of the WT (2.3 kcal/mol). These findings indicate that the mutational effect because of the induced fit on the FES landscape is remarkable, while the stabilization energies for both cases are small from a kinetic point of view. Because of the fluctuation along the d_2 direction as found in Fig. 7(c), there is room for a substrate to move in this pocket, resulting possibly in a decrease in the enzymatic activity. The present results clearly demonstrate that the FEL analyses followed by PaCS-MD is a powerful method for extracting the changes because of the site-directed mutation in the induced-fit processes as local but large conformational changes.

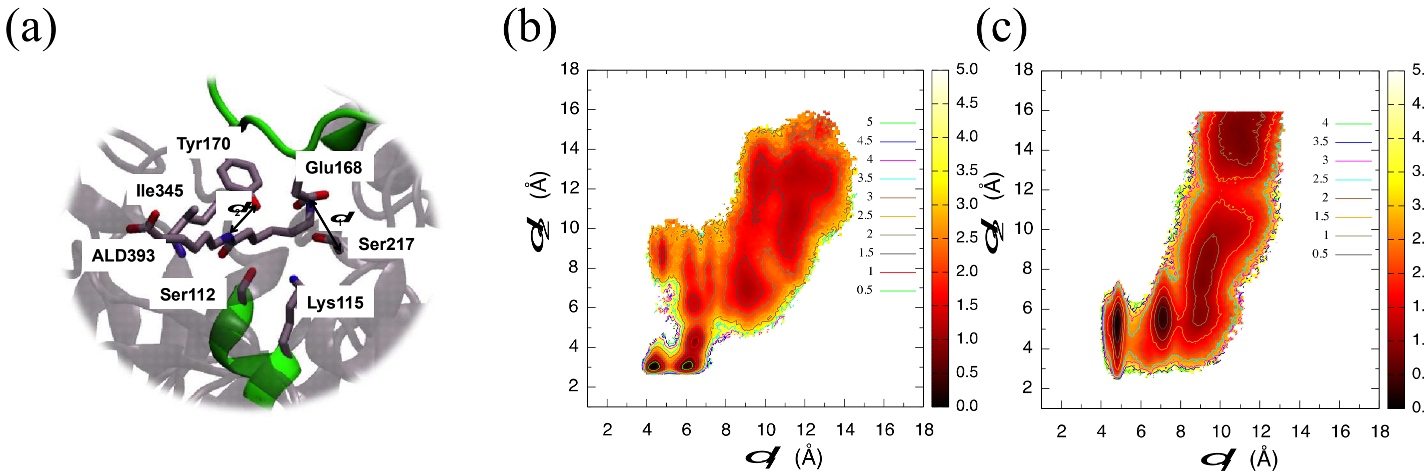


Figure 7. The local structures for the closed form (a). FEL of the WT (b) and the Y170F mutants (c) projected onto two collective variables, which are defined in (a) (in kcal/mol).†

3.3 Analyses of protein-folding pathways by TBSA and OFLOOD

The protein-folding processes are more difficult to extract compared with the collective domain motions of T4L and GlnBP treated in Sec. 3.1 because the conformational subspaces become broader and specifications of RCs to characterize the folding events are also non-trivial. If a set of the reactant and the product is given, *i.e.*, denatured and native structures are known *a priori*, PaCS-MD might provide their folding pathways. However, unfortunately, these situations are quite rare. On the other hand, FFM might provide the conformational transition pathway based on the anisotropic modes of proteins starting from a given reactant; however, the specifications for enhancing the PM might be difficult. Based on the above issues, as more powerful and general methods, TBSA and OFLOOD will be introduced through applications to the protein-folding simulations in terms of relatively small proteins.

For protein-folding pathways, let us consider the following miniproteins, Chignolin (PDBid: 1UAO)[88] and the villin headpiece subdomain called HP35 (PDBid: 1YRF),[89] which are designed to fold quickly to their native structures. Their folding pathways were extracted from completely extended structures using our methods, TBSA and OFLOOD. In both applications, the surrounding solvent was implicitly included using the generalized Born model (IGB = 5 in the SANDER module of AMBER 11)[69] with 0.2 M salt concentration and 0.005 kcal/(mol Å^2) surface tension. For the force fields used in the MD simulations, AMBER 99SB[70] and AMBER ff03[90] were selected for Chignolin and HP35, respectively, based on the preceding studies.[45, 91, 92]

3.3.1 Chignolin

As the first demonstration, folding pathways of Chignolin are extracted by TBSA. Chignolin is an artificial miniprotein consisting of 10 residues, amounting to 138 atoms, and forms a β-hairpin structure as the native structure because of hydrogen bonds in the backbone and hydrophobic packing between Y2 and W9 in the side chains.[88] As an initial structure, a completely stretched structure was employed based on the amino-acid sequences using the tLEaP module in AMBER 11.[69] In TBSA, states with low frequencies are intensively selected based on the inverse histogram projected onto the energy space and followed by conformational resampling from them. To obtain an initial histogram, 100 short-time (100-ps) MD simulations were

† Figure 7 has already been published in reference.[82] However, the layouts are different from the original layouts.

independently started from the completely stretched structure, where the initial velocities were differently assigned based on the Maxwell–Boltzmann distribution using the *NVT* ensemble (T = 200 K using the Berendsen thermostat, which is a stricter test with respect to simulations at 300 K to demonstrate the conformational sampling efficiency when TBSA is employed under such a low temperature environment). In terms of the obtained trajectories, an inverse histogram was calculated from the initial histogram, where they were used as weights for selecting the seeds with low frequencies in the energy space. To remove the dependence on the initial structure, 10 distinct trials (denoted as CH1, CH2, … , CH10) were independently performed. The cycle was repeated until the 10th cycle. In each cycle, the seeds were selected from the accumulated snapshots that belong to the energy bin, E, so that they are proportional to $\rho_{\mathrm{inv}}\left(E\right)$. In total, $\sum_{E=E_{\mathrm{min}}}^{E=E_{\mathrm{max}}} \rho_{\mathrm{inv}}\left(E\right)$ seeds were selected by summing the energy region [E_{min}, E_{max}]. The numbers of the selected seeds in each cycle for each trial are listed in Table 1 in reference.[58] For the convergence of the histogram with respect to the energy space, $-\ln\rho\left(E\right)$ was plotted, as shown in Fig. 8. In this figure, the distributions of the first three cycles (see the red, green, and blue lines in Fig. 8) drastically change, and the sampled energy region was gradually expanded towards the lower energy region. In contrast, those of the last two (the 9th and 10th) cycles coincide with each other (see the cyan and magenta lines in Fig. 8), indicating convergence of the conformational sampling in TBSA. As a comparison of the conformational efficiency of TBSA, a brute-force CMD simulation was started from the completely extended structure until 1 μs under the same *NVT* ensemble (T = 200 K). As Fig. 2(f) in reference[57] shows, the system rapidly folded into the misfolded state and remained trapped in the local minimum state, indicating the difficulty in complete protein folding despite the relatively long-time brute-force CMD. For the conformational sampling efficiency of TBSA, the minimum time to sample the native structure (C_{α} RMSD < 1.0 Å) was 16.1 ns (see Table 2 in reference),[58] which was shorter than the extremely long-time brute-force CMD by Shaw and co-workers.[3] Herein, their brute-force CMD simulations[4] needed 0.6 μs to sample the native structure of Chignolin. As another comparison of the sampling efficiency of TBSA, there is the McMD simulation with the same implicit model. In this McMD study,[92] a total of 180 ns multicanonical simulations were performed in addition to preliminary runs to obtain the DOS for constructing the multicanonical potential. The McMD simulation needed 50 ns (or more) to sample the native structure, indicating the improved conformational sampling efficiency of TBSA because our method needed only several nanosecond simulations without any preliminary run. As a comparison with recent studies,[52] MSES successfully sampled the native structure of Chignolin and described the protein-folding pathway by drawing its folding FEL. In MSES, to enhance the conformational sampling, coarse-grained and all-atom models are coupled with harmonic restraints. Because of the coupling, MSES enables one to perform a broad conformational sampling as a multi-scale simulation. To obtain the folding FEL, MSES needed a total of 0.6 μs simulation (HREM[54] using six replicas with 100 ns simulation for each replica for exchanging the harmonic restraints), highlighting the high conformational sampling efficiency of TBSA. It is stressed here that a remarkable advantage of TBSA over the above methods is that one does not need any preliminary run in applications. In TBSA, one only has to calculate and update *Inv*(E) to select seeds for the conformational resampling.

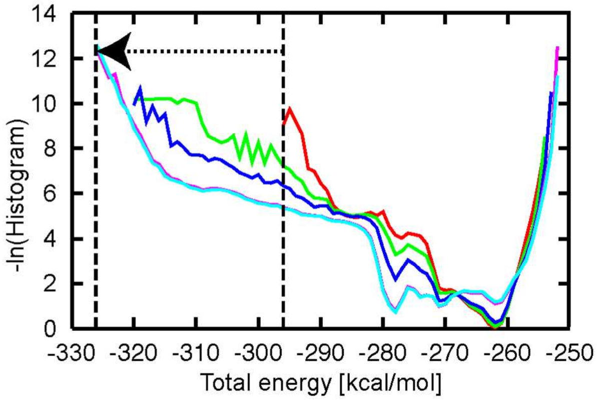


Figure 8. Profiles of histograms projected onto the total energy space. The red, green, and blue lines correspond to the negative logarithm of histograms at the first, second, and third cycles, respectively. The cyan and magenta lines correspond to those at the ninth and tenth cycles, respectively. The dashed arrow shows the explored region of total energy by TBSA during the 10 cycles.‡

As RCs for describing the folding energy landscape of Chignolin, the hydrogen-bonding distances in the backbone (HB1: D3(N)–G7(O), HB2: D3(N)–T8(O)) might be appropriate as adopted in the previous studies[45, 92] for the folding FEL analyses. To evaluate quantitatively the snapshots sampled by TBSA, these snapshots were projected onto the subspace spanned by HB1 and HB2 and their total energies were averaged in each projected point. Figure 9 shows the averaged total energy landscape, where the native and misfolded states were observed as local minima indicated by the arrows. As the landscape shows, stabilities of the structures sampled by TBSA were simply evaluated by averaged total energies without performing heavy FEL calculations, which might be an important strategy for obtaining a coarse-grained picture of energy landscapes.

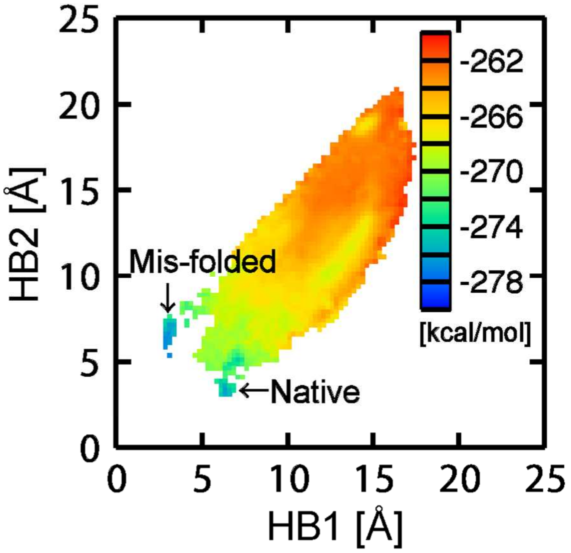


Figure 9. Total energy landscape of Chignolin projected onto the subspace spanned by HB1 and HB2 explored using TBSA. Each value of total energy was averaged over all snapshots in the projections. The native and misfolded states are indicated by arrows. The misfolded structure showed a good correspondence with the recent studies. In particular, the arrangements of the side chains (Y2 and W9) were the same (see Fig. 1 in reference),[93] which is supporting evidence for TBSA.§

‡ Figure 8 has already been published in reference.[58] However, the layout is different from the original layout.

§ Figure 9 has already been published in reference.[58] However, the layout is different from the original layout.

3.3.2 HP35

As a second application, the folding pathways of HP35 were extracted using OFLOOD. HP35 is a 35-residue miniprotein consisting of three bundle helices (Helix 1: residues 3–14, Helix 2: residues 15–21, and Helix 3: residues 22–33), amounting to 582 atoms.[89] In preceding studies,[44, 91] the orders of forming each helix in the folding processes have been discussed through the partial C_α RMSD defined by the following segments, segment A (residues 3–21, consisting of Helix 1–2) and segment B (residues 15–33, consisting of Helix 2–3). The overlap of Helix 2 in these segments ensures the overall folding of HP35 into the native structure once both segment A and segment B fold. Therefore, these partial C_α RMSDs might be appropriate RCs and they were specified as a set of RCs in OFLOOD for drawing two-dimensional distributions. As an initialization, short-time (100 ps × 100 runs) MD simulations were independently started from the completely extended structure with differently assigned initial velocities using the *NVT* (*T* = 300 K) ensemble. Herein, all MD simulations were performed using the SANDER module of AMBER 11.[69] Then, the distribution projected onto the subspace was calculated and followed by detection of outliers using the clustering algorithm FlexDice,[94-97] continuing to the conformational resampling from the detected outliers. Cycles of the detection of outliers and their conformational resampling were repeated until the 20th cycle by checking the accumulated distribution projected onto each RC (for the convergence of OFLOOD, see Fig. 3 in reference.[57] The total numbers of outliers detected during the 20 cycles were as follows: (99, 96, 82, 126, 73, 75, 66, 77, 101, 126, 79, 76, 88, 74, 66, 52, 40, 53, 61, and 65). For each seed, short-time (100-ps) MD simulation was performed by initializing velocities. Figure 10 shows the projections (red points) of the conformational resampling from the detected outliers (black points) every five cycles.

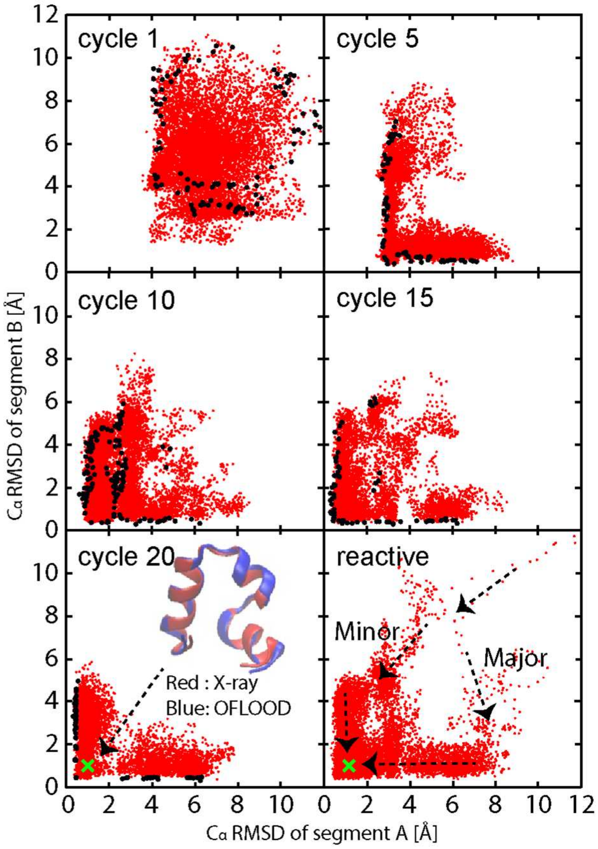


Figure 10. Projections of the snapshots generated by OFLOOD onto the subspace spanned by partial C_α RMSDs of each segment as red points. The black points correspond to the outliers detected by FlexDice in each

cycle. The cross (light green) corresponds to the projection of the X-ray structure of HP35. The structure with the minimum C_α RMSD (blue) is superimposed with the X-ray structure (red).[**]

Judging from these figures, it is clear that the detected outliers were located near the edge of the distributions and gradually expanded the distributions as the cycles went on. These facts numerically prove that our strategy to select the seeds and to perform resampling for these seeds works quite well for finding the folded structure. As shown in the middle right of Fig. 10, the native structure of HP35 was sufficiently sampled at the 15th cycle within the criterion (C_α RMSD < 1.0 Å). The results for 20 cycles together with the X-ray structure (black) are shown in the left bottom of Fig. 10 (the minimum overall C_α RMSD was 0.5 Å). Furthermore, the projections of the reactive trajectories were also projected onto the subspace in Fig. 10, denoted as “reactive”. From this figure, the projections might be split into two folding pathways, indicating the existence of a minor folding pathway in addition to the major folding pathway, which had been reported in preceding works.[44, 91] As experimental evidence, a recent triplet–triplet-energy transfer (TTET) experiment[98] supports the validity of the minor folding pathway extracted by OFLOOD. However, this minor folding pathway was not sampled by the brute-force 8-μs REMD simulations in the previous study,[91] showing the high conformational sampling efficiency of OFLOOD because OFLOOD has extracted such a rare event in sub-μs order computational time (total of 135.6 ns computational time for the 15 cycles). In contrast, a brute-force CMD simulation with Anton[3] needed 2.8 μs to sample the native structure from an unfolded structure. The above comparisons and discussions sufficiently demonstrate the effectiveness of TBSA in conformational searching in biomolecules.

3.4 Perspective and remaining problems

Here, we would like to mention future perspectives and remaining problems in our studies towards precise predictions for biologically relevant rare events. The first issue is related to a proper choice of RCs, which determines the conformational sampling efficiency. In our methods, the conformational sampling efficiency depends on selection of good seeds for promoting structural transitions of biomolecules from accumulated distributions projected upon characteristic RCs to describe biological reactions along them. This is one of the most serious remaining problems because the specifications of RCs are non-trivial issues in a wide variety of fields, and in general RCs strongly depend on the target systems. For instance, to extract collective domain motions of biomolecules, PCs defined by the PCA might be suitable because large-amplitude fluctuations upon the domain motions are low frequency modes and tend to be represented by several PCs projected onto a low-dimensional subspace spanned by a set of anisotropic PMs. On the other hand, allosteric effects upon ligand binding might not be represented by only the PMs because correlations among locally and globally fluctuating modes play essential roles in most of the allosteric functions. Therefore, the development of methodologies to extract the mode couplings among these modes and decompositions of collective modes into correlated and uncorrelated modes would be quite important issues to understand the allosteric functions. A full correlation analysis (FCA)[66] has partially solved this issue by setting “best axes” to characterize collective modes based on an independent component analysis (ICA).[99, 100] The basic concept of ICA/FCA is to select the best axes so as to regard a probability distribution projected onto these axes approximately as a product of Gaussians. For the mode couplings, an independent subspace analysis (ISA)[65] successfully extracted several independent subspaces from the conformational space that have significantly correlated collective modes in

[**] Figure 10 has already been published in the reference.[57] However, the layout is different from the original layout.

each subspace. In spite of these efforts, it is difficult to detect automatically correlations among the extracted subspaces. Thus, it is still necessary to develop an efficient protocol for automatically generating appropriate RCs, which might be useful to search relevant rare events together with our methods.

The second issue is related to how to evaluate precise structural transition pathways or networks based on our methods. To evaluate the transition pathways, our methods need some post-processing processes such as free energy calculations using the US or path-sampling methods, meaning that the coarse-grained transition pathways generated by our methods will be refined through well-established methods. Based on the above approaches, one can quantitatively evaluate the transition pathways, although relatively high computational costs might be required. Thus, it would be quite convenient to extract directly the transitional networks as kinetics without additional calculations. As an example of the analytical approach, the Markov State Model (MSM)[101-106] might be helpful for extracting directly the networks among states of biomolecules. In MSM, transitions among discretized states of biomolecules in a stationary state are estimated as a transition matrix, whose eigenvalues and eigenvectors describe probability densities and dynamical networks among the stationary states, respectively. Although this approach gives the transitional networks among microstates without the post-processing processes, the results of MSM strongly depend on how the discretized microstates are defined. In general, the discretization is performed by several clustering methods. Therefore, one needs robust clustering algorithms to define them and describe their networks as appropriately as possible.

The third issue is related to how to perform effectively the conformational sampling on the fly by referring only to past information. For instance, in the case of TBSA, the conformational sampling efficiency depends on an inverse histogram, which is derived from the past conformational sampling, to sample efficiently states with low frequencies for selecting seeds. Because of the conformational resampling based on the inverse histogram, the conformational spaces are updated on the fly, ideally achieving converged spaces, as demonstrated in Fig. 8. However, a free energy surface is not simultaneously obtained in the current scheme, although a potential energy surface averaged over generated snapshots is available. To calculate the energy surface on the fly without any post-processing, the current scheme needs some extensions. As an example of on-the-fly schemes for the free energy calculations, mean force dynamics simultaneously estimate free energies as time-dependent energies based on a conserved quantity,[107] meaning that the FELs are updated on the fly, and ideally they would converge after long-time simulations. For instance, the recently developed logarithmic mean-force dynamics (LogMFD)[108, 109] successfully calculates the FELs without any additional cost. LogMFD refers to logarithmic forms of free energies as additional potentials to flatten free energy barriers among minima. As a future extension of our methods, this type of self-updating scheme might be applied to evaluate free energies on the fly along with estimations of transition pathways.

The fourth issue is how to know when products have been reached in the different sampling methods. For this issue, we think that there are two different situations: (i) a set of reactants and products is known *a priori*, and (ii) only reactants are known. For the first situation, it might be relatively easy to know when the target has been reached to a desirable conformation because the conformational transitions can be monitored based on the RMSD measured from the products, as used in the PaCS-MD method.[55] For instance, in the previous work on PaCS-MD, we showed that RMSD gradually converges to small values, say 1.0 Å is enough for the conformational search, with an increase in the PaCS-MD cycles. On the other hand, for the second situation, it might not be easy to judge when the targets have reached native and/or metastable structures in the different methods. As a strategy for this judgement, we have already proposed a criterion for each method based on the convergence of distributions projected onto a set of RCs and explained the scheme in our previous studies. For instance, in TBSA, Eq. (2) corresponds to the criterion. If this quantity converges to a sufficiently small value, the conformational searches by TBSA are well accomplished. Of course, these types of criteria should be appropriately defined depending on the method employed and checked for convergence in the actual calculations, as we demonstrated. The criteria both for FFM and OFLOOD methods have also been concisely explained in the original papers.[56, 57]

The fifth issue is whether or not the different conformational sampling methods provide reasonable ensembles. Herein, the ensembles obtained from our methods might be slightly different from true canonical ensembles. Patched trajectories obtained from distinct short-time MD simulations, which we referred to as “reactive trajectories”, are generated without any reweighting. However, these reactive trajectories might be helpful for constructing fine FELs with the help of well-established free energy calculation methods such as US because the reactive trajectories might give a coarse-grained ensemble, as shown in Figs. 3 and 4. However, as a future perspective of this study, direct refinements of the reactive trajectories with reasonable reweighting techniques and without any post-processing calculations are needed.

4. Conclusions

Enhanced conformational sampling methods are indispensable for extracting the conformational transitions of biomolecules because they are strongly related to the biological functions. Because of the problems of accessible timescales in CMD simulations and stochastic processes of the biologically relevant rare events, the conformational transitions relevant to biological functions are quite difficult to extract, even if brute-force CMD simulations are employed using current massively parallel computational resources. To tackle these types of problems, we would like to propose several enhanced conformational sampling methods based on the conformational resampling through short-time MD simulations. In our strategy, the seeds that correspond to initial structures for short-time MD simulations are first selected. Here, the seeds should be selected so that they have the potential to transit to other metastable states with high probability. Next, multiple short-time MD simulations are independently restarted from the selected seeds via initialization of velocities. Based on this strategy, each method was applied for the purpose of studies to extract biologically rare events of proteins under the following conditions: (i) a set of reactant and product is given *a priori* and (ii) only a single reactant is given.

For the first case, PaCS-MD is applicable to extract biologically rare events such as structural transitions of proteins that connect the reactant to the product. As examples of structural transitions without ligand bindings, the open–closed domain motions of T4L/GlnBP were extracted with ns-order simulations by PaCS-MD. For a more complicated system such as the induced-fit process of the WT and the Y170F mutant of NylB (enzyme) to Ald (substrate), although the FEL changes drastically with respect to the WT case, the free energy barrier is nearly unchanged, meaning that the kinetics of the induced fit are unaffected by the mutation or, at least, this mutation. As shown in these applications, PaCS-MD is a powerful method to extract structural transitions of proteins with/without ligand binding. For the second case, FFM, TBSA, and OFLOOD are applicable to extract transition pathways starting from a given reactant. For instance, FFM derived the open–closed domain motions of T4L starting from the open structure with ns-order simulations. TBSA and OFLOOD also extracted several folding pathways of miniproteins with ns-order simulations starting from the completely extended structures. These types of methods might be convenient to predict transition pathways or find local/global energy minimum states of proteins starting from a given reactant without knowing the reactants, which is one of the advantages of conformational sampling and indicates the generality of our methods.

As future perspectives, we discussed the following three issues to be solved for further enhancements of conformational sampling. (i) The automatic specification of RCs. (ii) The refinements of transition pathways and networks generated by each method. (iii) Extensions to the on-the-fly algorithm for the free energy calculations, which has already been discussed in Sec. 3.4.

Finally, we would like to mention the implementation of our methods. From the point of view of implementation, our methods can be easily implemented in MD software by using script-like procedures without modifying any source codes. Therefore, any MD software packages can be used for applications.

ACKNOWLEDGMENTS

This work is supported by Grants-in-Aid for Scientific Research of the Innovative Areas “Computics”, “Dynamic ordering and biofunction”, “Photosynergetics”, and “3-D Active-Site Science” (No. 25104716, 25104717, 26102525, 26107004, and 26105012) from the Japan Society for the Promotion of Science (JSPS), and by the Strategic Programs for Innovative Research (SPIRE), MEXT, and the Computational Materials Science Initiative (CMSI), Japan. This research is also supported by a Grant-in-Aid for Young Scientists (B) (No. 26840056) from JSPS KAKENHI. TB expresses his special thanks for the JSPS Research Fellowship for Young Scientists (No. 02503029). Theoretical calculations were partly performed at the Research Centre for Computational Science, Institute for Molecular Science (Japan) and the Centre of Computational Materials Science, Institute for Solid State Physics, The University of Tokyo (Japan).

References

1. R. Harada, N. Tochio, T. Kigawa, Y. Sugita and M. Feig, *J. Am. Chem. Soc.*, 2013, **135**, 3696.
2. R. Harada, Y. Sugita and M. Feig, *J. Am. Chem. Soc.*, 2012, **134**, 4842.
3. K. Lindorff-Larsen, S. Piana, R. O. Dror and D. E. Shaw, *Science*, 2011, **334**, 517.
4. D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. B. Shan and W. Wriggers, *Science*, 2010, **330**, 341.
5. S. Piana, K. Lindorff-Larsen and D. E. Shaw, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 5915.
6. E. Elsen, M. Houston, V. Vishal, E. Darve, P. Hanrahan and V. Pandee, *Proceddings of the 2006 ACM/IEEE Conference on Supercomputing. New York, NY, 2006*, 2006.
7. T. Mashimo, Y. Fukunishi, N. Kamiya, Y. Takano, I. Fukuda and H. Nakamura, *J. Chem. Theory Comput.*, 2013, **9**, 5599.
8. N. Kamiya, I. Fukuda and H. Nakamura, *Chem. Phys. Lett.*, 2013, **568**, 26.
9. I. Fukuda, N. Kamiya, Y. Yonezawa and H. Nakamura, *J. Chem. Phys.*, 2012, **137**.
10. I. Fukuda, Y. Yonezawa and H. Nakamura, *J. Chem. Phys.*, 2011, **134**.
11. A. W. Gotz, M. J. Williamson, D. Xu, D. Poole, S. Le Grand and R. C. Walker, *J. Chem. Theory Comput.*, 2012, **8**, 1542.
12. R. Salomon-Ferrer, A. W. Gotz, D. Poole, S. Le Grand and R. C. Walker, *J. Chem. Theory Comput.*, 2013, **9**, 3878.
13. N. Goga, S. Marrink, R. Cioromela and F. Moldoveanu, *Proccedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus*, 2012.
14. J. C. Phillips, J. E. Stone and K. Schulten, *Proceedings of the 2008 ACM/IEEE Conference on Supercomputing, Piscataway, NJ*, 2008.
15. P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L. P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts and V. S. Pande, *J. Chem. Theory Comput.*, 2013, **9**, 461.
16. C. Dellago, P. G. Bolhuis, F. S. Csajka and D. Chandler, *J. Chem. Phys.*, 1998, **108**, 1964.
17. P. G. Bolhuis, D. Chandler, C. Dellago and P. L. Geissler, *Annu. Rev. Phys. Chem.*, 2002, **53**, 291.
18. C. Dellago, P. G. Bolhuis and P. L. Geissler, *Advances in Chemical Physics, Vol 123*, 2002, **123**, 1.
19. E. Weinan, W. Q. Ren and E. Vanden-Eijnden, *Phys Rev B*, 2002, **66**.
20. E. Weinan, W. Q. Ren and E. Vanden-Eijnden, *J..Phys. Chem. B*, 2005, **109**, 6688.
21. L. Maragliano, A. Fischer, E. Vanden-Eijnden and G. Ciccotti, *J. Chem. Phys.*, 2006, **125**, 024106.
22. L. Maragliano and E. Vanden-Eijnden, *Chem. Phys. Lett.*, 2007, **446**, 182.
23. E. Vanden-Eijnden and M. Venturoli, *J. Chem. Phys.*, 2009, **130**.
24. Y. Matsunaga, H. Fujisaki, T. Terada, T. Furuta, K. Moritsugu and A. Kidera, *PLoS Comput. Biol.*, 2012, **8**.
25. J. Schlitter, M. Engels, P. Kruger, E. Jacoby and A. Wollmer, *Mol. Simulat*, 1993, **10**, 291.
26. B. Isralewitz, J. Baudry, J. Gullingsrud, D. Kosztin and K. Schulten, *J. Mol. Graph. Model.*, 2001, **19**, 13.
27. B. Isralewitz, M. Gao and K. Schulten, *Curr. Opin. Struct. Biol.*, 2001, **11**, 224.
28. A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 12562.
29. M. Iannuzzi, A. Laio and M. Parrinello, *Phys. Rev. Lett.*, 2003, **90**.
30. A. Laio and F. L. Gervasio, *Rep. Prog. Phys.* , 2008, **71**, 126601.
31. F. L. Gervasio, M. Boero and M. Parrinello, *Angew. Chem. Int. Edit.*, 2006, **45**, 5606.
32. M. Boero, *J. Phys.Chem. B*, 2011, **115**, 12276.
33. X. Biarnes, A. Ardevol, J. Iglesias-Fernandez, A. Planas and C. Rovira, *J. Am. Chem. Soc.*, 2011, **133**, 20301.
34. M. Boero, T. Ikeda, E. Ito and K. Terakura, *J. Am. Chem. Soc.*, 2006, **128**, 16798.
35. K. Kamiya, M. Boero, M. Tateno, K. Shiraishi and A. Oshiyama, *J. Am. Chem. Soc.*, 2007, **129**, 9663.
36. Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.*, 1999, **314**, 141.
37. U. H. E. Hansmann, Y. Okamoto and F. Eisenmenger, *Chem. Phys. Lett.*, 1996, **259**, 321.

38. N. Nakajima, H. Nakamura and A. Kidera, *J. Phys. Chem. B*, 1997, **101**, 817.
39. J. Higo, K. Umezawa and H. Nakamura, *J. Chem. Phys.*, 2013, **138**.
40. J. Ikebe, S. Sakuraba and H. Kono, *J. Comput. Chem.*, 2014, **35**, 39.
41. P. Liu, B. Kim, R. A. Friesner and B. J. Berne, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 13749.
42. H. Okumura, *J. Chem. Phys.*, 2008, **129**.
43. Y. Okamoto, *J. Mol. Graph. Model.*, 2004, **22**, 425.
44. R. Harada and A. Kitao, *J. Chem. Theory Comput.*, 2012, **8**, 290.
45. R. Harada and A. Kitao, *J. Phys. Chem. B*, 2011, **115**, 8806.
46. R. Harada and A. Kitao, *Chem. Phys. Lett.*, 2011, **516**, 113.
47. G. M. Torrie and J. P. Valleau, *J Comput Phys*, 1977, **23**, 187.
48. G. M. Torrie and J. P. Valleau, *Chem. Phys. Lett.*, 1974, **28**, 578.
49. M. Souaille and B. Roux, *Comput Phys Commun*, 2001, **135**, 40.
50. S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman and J. M. Rosenberg, *J. Comput. Chem.*, 1992, **13**, 1011.
51. A. M. Ferrenberg and R. H. Swendsen, *Phys. Rev. Lett.*, 1989, **63**, 1195.
52. K. Moritsugu, T. Terada and A. Kidera, *J. Chem. Phys.*, 2010, **133**, 224105.
53. Y. Sugita, A. Kitao and Y. Okamoto, *J. Chem. Phys.*, 2000, **113**, 6042.
54. H. Fukunishi, O. Watanabe and S. Takada, *J. Chem. Phys.*, 2002, **116**, 9058.
55. R. Harada and A. Kitao, *J. Chem. Phys.*, 2013, **139**, 035103.
56. R. Harada, Y. Takano and Y. Shigeta, *J. Chem. Phys.*, 2014, **140**, 125103
57. R. Harada, T. Nakamura, Y. Takano and Y. Shigeta, *J. Comput. Chem.*, 2015, **36**, 97
58. R. Harada, Y. Takano and Y. Shigeta, *J. Comput. Chem., in press*.
59. L. S. D. Caves, J. D. Evanseck and M. Karplus, *Protein Sci.*, 1998, **7**, 649.
60. D. Bhatt, B. W. Zhang and D. M. Zuckerman, *J. Chem. Phys.*, 2010, **133**, 014110.
61. M. Nakano, H. Watanabe, S. M. Rothstein and S. Tanaka, *J. Phys. Chem. B*, 2010, **114**, 10234.
62. A. Amadei, A. B. Linssen and H. J. Berendsen, *Proteins*, 1993, **17**, 412.
63. A. Kitao, F. Hirata and N. Go, *Chem. Phys.*, 1991, **158**, 447.
64. A. Kitao, *J. Chem. Phys.*, 2011, **135**, 045101.
65. S. Sakuraba, Y. Joti and A. Kitao, *J. Chem. Phys.*, 2010, **133**, 185102.
66. O. F. Lange and H. Grubmuller, *Proteins-Structure Function and Bioinformatics*, 2008, **70**, 1294.
67. L. H. Weaver and B. W. Matthews, *J. Mol. Biol.*, 1987, **193**, 189.
68. X. J. Zhang and B. W. Matthews, *Protein Sci.*, 1994, **3**, 1031.
69. T. A. D. D. A. Case, T. E. Cheatham, III, C. L. Simmerling, J. Wang, R. E. Duke, R., R. C. W. Luo, W. Zhang, K. M. Merz, B. Roberts, B. Wang, S. Hayik, A. Roitberg, I. K. G. Seabra, K. F. Wong, F. Paesani, J. Vanicek, J. Liu, X. Wu, S. R. Brozell, H. G. T. Steinbrecher, Q. Cai, X. Ye, J. Wang, M.- J. Hsieh, G. Cui, D. R. Roe, D. H., M. G. S. Mathews, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, P. A. Kollman, AMBER 11; University of California: San Francisco, 2010.
70. V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins: Struct., Funct., Genet.*, 2006, **65**, 712.
71. H. J. C. Berendsen, J. P. M. Postma, W. F. Vangunsteren, A. Dinola and J. R. Haak, *J. Chem. Phys.*, 1984, **81**, 3684.
72. C. D. Hsiao, Y. J. Sun, J. Rose and B. C. Wang, *J. Mol. Biol.*, 1996, **262**, 225.
73. Y. J. Sun, J. Rose, B. C. Wang and C. D. Hsiao, *J. Mol. Biol.*, 1998, **278**, 219.
74. D. A. Silva, G. R. Bowman, A. Sosa-Peinado and X. H. Huang, *PLoS Comput. Biol.*, 2011, **7**.
75. H. X. Kondo, N. Okimoto, G. Morimoto and M. Taiji, *J. Phys. Chem. B*, 2011, **115**, 7629.
76. J. H. Peng and Z. Y. Zhang, *J. Chem. Theory Comput.*, 2014, **10**, 3449.
77. H. Okada, S. Negoro, H. Kimura and S. Nakamura, *Nature*, 1983, **306**, 203.
78. S. Negoro, *Biopolymers*, 2002, **9**, 395.
79. T. Baba, K. Kamiya, T. Matsui, N. Shibata, Y. Higuchi, T. Kobayashi, S. Negoro and Y. Shigeta, *Chem. Phys. Lett.*, 2011, **507**, 157.
80. H. Ando, Y. Shigeta, T. Baba, C. Watanabe, Y. Okiyama, Y. Mochizuki and M. Nakano, *Mol. Phys.*, 2014,

DOI: Doi 10.1080/00268976.2014.941311.

81. K. Kamiya, T. Baba, M. Boero, T. Matsui, S. Negoro and Y. Shigeta, *J. Phys. Chem. Lett.*, 2014, **5**, 1210.
82. T. Baba, R. Harada, M. Nakano and Y. Shigeta, *J. Comput. Chem.*, 2014, **35**, 1240.
83. T. Baba, M. Boero, K. Kamiya, H. Ando, S. Negoro, M. Nakano and Y. Shigeta, *Phys. Chem. Chem. Phys.*, 2014.
84. S. Negoro, T. Ohki, N. Shibata, K. Sasa, H. Hayashi, H. Nakano, K. Yasuhira, D. I. Kato, M. Takeo and Y. Higuchi, *J. Mol. Biol.*, 2007, **370**, 142.
85. Y. Kawashima, T. Ohki, N. Shibata, Y. Higuchi, Y. Wakitani, Y. Matsuura, Y. Nakata, M. Takeo, D. Kato and S. Negoro, *FEBS J.*, 2009, **276**, 2547.
86. T. A. D. D. A. Case, T. E. Cheatham, III, C. L. Simmerling, J. Wang, R. E. Duke, R., R. C. W. Luo, W. Zhang, K. M. Merz, B. Roberts, B. Wang, S. Hayik, A. Roitberg, I. K. G. Seabra, K. F. Wong, F. Paesani, J. Vanicek, J. Liu, X. Wu, S. R. Brozell, H. G. T. Steinbrecher, Q. Cai, X. Ye, J. Wang, M.- J. Hsieh, G. Cui, D. R. Roe, D. H., M. G. S. Mathews, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, P. A. Kollman, AMBER 12; University of California: San Francisco, 2012.
87. J. M. Wang, P. Cieplak and P. A. Kollman, *J. Comput. Chem.*, 2000, **21**, 1049.
88. S. Honda, K. Yamasaki, Y. Sawada and H. Morii, *Structure*, 2004, **12**, 1507.
89. T. K. Chiu, J. Kubelka, R. Herbst-Irmer, W. A. Eaton, J. Hofrichter and D. R. Davies, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 7517.
90. Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. M. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. M. Wang and P. Kollman, *J. Comput. Chem.*, 2003, **24**, 1999.
91. H. X. Lei, C. Wu, H. G. Liu and Y. Duan, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 4925.
92. D. Satoh, K. Shimizu, S. Nakamura and T. Terada, *FEBS Lett.*, 2006, **580**, 3422.
93. P. Kuhrova, A. De Simone, M. Otyepka and R. B. Best, *Biophys. J.*, 2012, **102**, 1897.
94. T. Nakamura, Y. Kamidori, S. Wakabayashi and N. Yoshida, *IPSJ Journal Database*, 2005, **46**, 40.
95. T. Nakamura, Y. Kamidoi, S. Wakabayashi and N. Yoshida, *IEEE Computer Press. SWOD2006*, 2006.
96. T. Nakamura, Y. Kamidoi, S. Wakabayashi and N. Yoshida, *IEEE Computer Press. Proc. on RIDE-SDMA2005*, 2005.
97. T. Nakamura, Y. Kamidoi, S. Wakabayashi and N. Yoshida, *IEEE Computer Press. Proc. on SWOD2005*, 2005.
98. A. Reiner, P. Henklein and T. Kiefhaber, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 4955.
99. A. M. Fraser and H. L. Swinney, *Phys. Rev. A*, 1986, **33**, 1134.
100. P. Comon, *Signal Process*, 1994, **36**, 287.
101. J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill and W. C. Swope, *J. Chem. Phys.*, 2007, **126**.
102. A. C. Pan and B. Roux, *J. Chem. Phys.*, 2008, **129**.
103. J. H. Prinz, J. D. Chodera, V. S. Pande, W. C. Swope, J. C. Smith and F. Noe, *J. Chem. Phys.*, 2011, **134**.
104. J. H. Prinz, B. Keller and F. Noe, *Phys. Chem. Chem. Phys.*, 2011, **13**, 16912.
105. J. H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schutte and F. Noe, *J. Chem. Phys.*, 2011, **134**.
106. J. D. Chodera and F. Noe, *Curr. Opin. Struct. Biol.*, 2014, **25**, 135.
107. L. Rosso, P. Minary, Z. W. Zhu and M. E. Tuckerman, *J. Chem. Phys.*, 2002, **116**, 4389.
108. T. Morishita, S. G. Itoh, H. Okumura and M. Mikami, *Phys. Rev. E*, 2012, **85**.
109. T. Morishita, S. G. Itoh, H. Okumura and M. Mikami, *J. Comput. Chem.*, 2013, **34**, 1375.