A novel algorithm for enhanced conformational sampling and its applications to compute free-energy landscapes of protein–protein and protein–ligand interactions

Tomonori Hayami
Graduate School of Frontier Biosciences
Osaka University

March 2023

Abstract

It is known that there is relationship between protein conformation and function, and various experiments and computer simulations have been used to study the relationship. In recent years, computer-aided conformational sampling is becoming increasingly important to investigate biomolecular phenomena including protein–ligand, protein–protein, and protein–DNA binding processes as well as protein folding. In many cases, generalized-ensemble methods are used to enhance sampling. For example, adaptive umbrella sampling, apply an effective potential, which is derived from temporarily assumed canonical distribution as a function of one or more arbitrarily defined reaction coordinates. However, it is not straightforward to estimate the appropriate canonical distribution, especially for cases applying multiple reaction coordinates. Multidimensional virtual-system coupled canonical molecular dynamics (mD-VcMD), which is one of the generalized-ensemble MD methods does not rely on the form of the canonical distribution. Therefore, it is practically useful to explore a high-dimensional reaction-coordinate space.

In this study, I applied the mD-VcMD method to two types of system to verify its usefulness in the field of bioscience. At first, I performed the method with the simple molecular models consisting of three or four alanine peptides. I confirmed that mD-VcMD efficiently searched 2D and 3D reaction-coordinate spaces defined as inter-peptide distances.

Next, I applied the method to three systems consisting of mSin3B and one of three compounds, sertraline, YN3, and acitretin. Sertraline, YN3, and acitretin are chemical compounds designed to inhibit binding of neural restrictive silencer factor/RE1-silencing transcription factor (NRSF/REST) to a corepressor mSin3B. These compounds can be a drug candidate for neurological diseases, such as Down's syndrome, medulloblastoma, Huntington disease, cardiomyopathy, and neuropathic pain. The mD-VcMD method produced useful quantities such as the spatial density of the ligand around the receptor, the intermolecular contact patterns, the propensity of molecular orientation, and the ligand flexibility. From these analyses, I showed that only sertraline produces a similar inter-molecular binding mode observed in the REST/NRSF–mSin3B complex.

Contents

Chapter I

General Introduction

Proteins are functional molecules with various roles in biological entities. Proteins are polypeptide chains with 20 types of amino acids as basic structural units, and form three-dimensional structures by folding from one or more polypeptide chains. As seen in the Anfinsen's research, proteins exerted their functions by spontaneously folding into a thermodynamically stable conformation called the native state. It was once thought that the stable conformation was determined only by the amino acid sequence. However, subsequent studies revealed that proteins form aggregates like amyloid within the cell and loses its function, which can be the cause of disease. In addition, a protein "chaperone" was discovered that suppresses the formation of these aggregates in cells and induces them to fold into the correct three-dimensional conformation. Furthermore, the discovery of intrinsically disordered proteins or intrinsically disordered regions that play various roles without adopting a specific three-dimensional structure in cells led to a revision of the above idea, known as Anfinsen's dogma.

The three-dimensional structures of proteins have been elucidated by experimental techniques such as X-ray crystallography, NMR, and cryo-electron microscopy. In addition to these experimental techniques, computer simulation has been also used to reveal the relationships between structures and functions. In this study, computer simulations of proteins are performed using, as it is called, the molecular dynamics method based on classical mechanics. But I utilize several techniques for efficient computation. If the number of atoms to be calculated is N, molecular dynamics simulation requires a computational load on the order of N^2, and even with a supercomputer, a huge amount of computational time is required to obtain useful results. Therefore, in order to discover stable conformations among possible protein conformations, or to discover stable conformations in protein–ligand interactions, computational methods to efficiently sample the structural spaces are required. To tackle this problem, the generalized-ensemble algorithms such as the multicanonical molecular dynamics method and the replica exchange method have been developed.

Multidimensional virtual-system coupled canonical MD (mD-VcMD) method used in this research is also positioned as one of these generalized-ensemble methods. In this mD-VcMD method, the internal parameters related to the structure of the protein in the system to be calculated, or the protein–protein/protein–ligand distances are set as reaction coordinates. And efficient samplings are performed along these reaction coordinates. Therefore, it is very useful for detailed investigation of proteins with flexible steric structures such as intrinsically disordered proteins, and interactions between proteins or between proteins and ligands.

In this study, I performed computer simulations using the mD-VcMD method to examine interactions between multiple peptides, and the interactions between a protein with a flexible three-dimensional structure called mSin3B and compounds that can bind to it. I demonstrated that the mD-VcMD method satisfies both the efficiency and accuracy requirements, and that it is possible to explore and sample the conformational space to investigate the intermolecular interactions in detail. The result of this research shows the advantages to investigate a target that is difficult to experimentally clarify the relationship between structure and function due to its flexible three-dimensional structure, such as intrinsically disordered proteins. I believe that it will be useful for the progress in this field in the future.

The mD-VcMD method is explained in Chapter II, the results of applying this method to simple systems containing 3 and 4 molecules of alanine-peptide are described in Chapter III, and the results are applied to the systems containing mSin3B and its ligand molecules in Chapter IV.

Chapter II

Methods

2.1 Introduction

In recent years, computer-aided conformational sampling is becoming increasingly important to investigate biomolecular phenomena including protein-ligand, protein-protein, and protein-DNA binding processes as well as protein folding. Particularly, generalized-ensemble methods that enhance sampling along a parameter defined by the configuration of a system, or a reaction coordinate, have analyzed the free-energy landscapes of such biomolecular systems. A major class of methods uses energy (or entropy) as a reaction coordinate and enhances conformational changes by scaling energy[1–20]. This energy-enhancement approach is powerful and suitable for identifying major energy basins, to which a high probability of existence is partitioned, in a high-dimensional and complicated potential energy space[21,22]. However, if a less-stable minor basin overlaps to the major basins in the energy axis, then the minor one is overlooked. If the minor basin has importance in the study, then the oversight is an important shortcoming[23–25]. Because this oversight is a natural outcome of energy-enhancement sampling, another type of reaction coordinate is required to search minor basins.

An alternative class of generalized-ensemble methods including umbrella sampling[26,27], adaptive umbrella sampling (AUS)[28–31], and their variants[13,32–38] introduces a structural parameter, for example, intermolecular distance, for the reaction coordinate. In theory, AUS can avoid the oversight described above when major and minor basins can be discriminated on the reaction-coordinate axis. To drive conformational changes along the reaction-coordinate axis, AUS applies an effective potential E_{AUS} as

$$\boldsymbol{E_{AUS} = E_R + RT\ln[P_{cano}(\lambda; T)]}, \tag{2.1}$$

where E_R represents the original potential energy defined by a force field, $P_{cano}(\lambda; T)$ denotes a canonical distribution function along the reaction coordinate $\lambda(\boldsymbol{r})$ at temperature T, and R denotes the gas constant. The reaction coordinate λ, also denoted as $\lambda(\boldsymbol{r})$, is a function of atomic coordinates of the system, $\boldsymbol{r} = [x_1, y_1, z_1, x_2, y_2, z_2, \cdots, x_N, y_N, z_N]$, where x_i, y_i, and z_i, respectively, denote the x-, y-, and z-coordinates of the i-th element in the Cartesian space, and N represents the number of atoms in the system. For simplicity, I abbreviate arguments $\boldsymbol{r}$ and T, respectively, for functions $\lambda(\boldsymbol{r})$ and $P_{cano}(\lambda; T)$, hereinafter. Molecular dynamics (MD) and Monte Carlo (MC) sampling using E_{AUS} instead of E_R generates a conformational ensemble with a uniform probability of existence along the reaction-coordinate axis. To apply AUS for studying molecular phenomena, the phenomena must be well characterized by conformational changes along with the reaction-coordinate axis. In other words, users must find an appropriate definition of the reaction coordinate according to their interest. In general, phenomena in biomolecular systems tend to be highly complicated. A single structural parameter is insufficient to characterize the conformational changes[25]. To overcome this, some methods using multiple reaction coordinates have been proposed[39–42], where sampling is achieved in the multidimensional reaction-coordinate space. In multidimensional AUS, eq. (2.1) is replaced with the following equation as

$$\boldsymbol{E_{AUS}} = \boldsymbol{E_R} + \boldsymbol{RT}\,\mathbf{ln}[\boldsymbol{P_{cano}}(\boldsymbol{\lambda})], \tag{2.2}$$

where $\boldsymbol{\lambda}$ is a vector of reaction coordinates, $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \cdots, \lambda_{N_{Rc}}]$, and N_{Rc} denotes the number of reaction coordinates. The multidimensional AUS can analyze the detailed free-energy landscape of complicated molecular phenomena in theory.

However, AUS requires precise estimation of the canonical distribution $P_{cano}(\boldsymbol{\lambda})$ as a function of reaction coordinates to calculate the effective potential at fixed T. Because $P_{cano}(\boldsymbol{\lambda})$ is unknown a priori in general, AUS simulation should be begun with an initial guess of $P_{cano}(\boldsymbol{\lambda})$. The guess is improved through iterative AUS simulations until a resultant ensemble uniformly distributes along λ. For this procedure, it is not guaranteed to converge the distribution. Accurate estimation of the rugged differentiable function $P_{cano}(\boldsymbol{\lambda})$ would be practically problematic with the use of a limited number of samples. Particularly, applying a higher-dimensional reaction-coordinate space requires a higher-dimensional differentiable function for the canonical distribution, $P_{cano}(\boldsymbol{\lambda}) = P_{cano}(\lambda_1, \lambda_2, \lambda_3, \cdots)$ at fixed T, which makes the estimation more difficult because of the so-called "curse of dimensionality." Therefore, development of a new method enhancing the conformational sampling in a high-dimensional reaction-coordinate space without explicit estimation of high-dimensional canonical distribution as a continuous function is highly anticipated.

Here, I introduce a new generalized-ensemble approach termed multidimensional virtual-system coupled canonical MD (mD-VcMD) to tackle this issue. This method is based on a series of virtual-system coupled sampling methods developed by Higo et al. These methods introduce a nonphysical system, called a virtual system, interacting with the physical system (molecular system or real system) to enhance conformational sampling[43]. The virtual system consists of some discrete states or virtual states. Transitions between them control the effective potential in the real system. In the real system, conformational changes are facilitated by interaction with the virtual system. Recently, Higo et al. presented the virtual system-coupled canonical sampling method and applied it for MC sampler with single and multiple reaction-coordinate systems, respectively, denoted as 1D-VcMC[24] and mD-VcMC[25]. In addition to making the sampling applicable to more realistic molecular systems, which are expressed by all-atom models in explicit solvent, I recently reported an MD version of this approach with a single reaction coordinate, termed VcMD[43]. This method requires no estimation of the explicit form of the canonical distribution function.

Because enhancement along a single reaction coordinate is sometimes insufficient for complex molecular systems, as described above, here I introduce an extension of VcMD to explore multiple reaction-coordinate space, termed mD-VcMD.

2.2 Entire system of mD-VcMD

As described in the "2.1 Introduction" section, mD-VcMD introduces a nonphysical system or virtual system, to enhance conformational sampling in the physical system or the real system. The coordinate of the virtual system is expressed as an integer vector, $\boldsymbol{L} = \left[L_1, L_2, \cdots, L_{N_{Rc}}\right]$, and migration of the coordinate $\boldsymbol{L}$ enhances conformational motions (CFMs) along $\boldsymbol{\lambda}$. Details of coupling between $\boldsymbol{L}$ and $\boldsymbol{\lambda}$ are described later. Because the entire system of mD-VcMD comprises the real and virtual systems, the phase space of mD-VcMD is defined formally as

$$\boldsymbol{\phi} = [\boldsymbol{r}, \boldsymbol{v}, \boldsymbol{L}], \tag{2.3}$$

where $\boldsymbol{v}$ denotes a $3N$-dimensional vector describing atomic velocities. Each discrete position in the virtual system is a virtual state. The number of discrete states in each of the N_{Rc} axes is defined arbitrarily by users (Fig. 2-1). Migration of the real and virtual coordinates is performed independently based, respectively, on the equation of motion and MC method, as explained later.

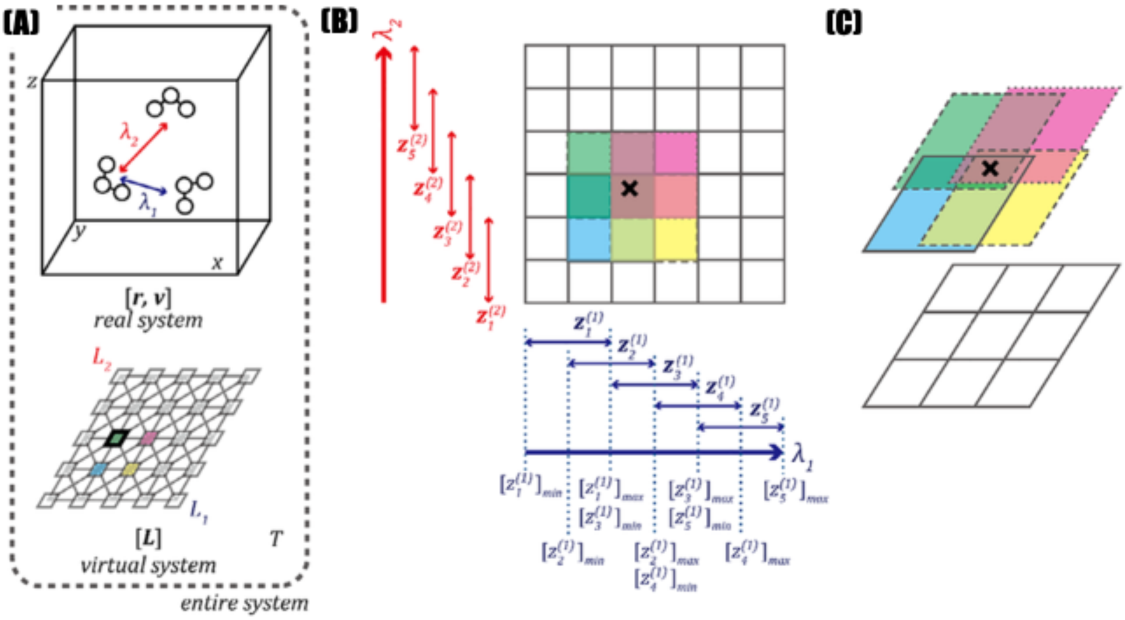

Figure 2-1. Schematic illustration of multidimensional virtual-system coupled canonical molecular dynamics (mD-VcMD). This figure shows the case for two-dimensional (2D)-VcMD as an example, where two reaction coordinates, λ_1 and λ_2, as well as two virtual coordinates L_1 and L_2, are introduced as explained below. (A) The entire system composed of the real system and virtual system. The former is a usual molecular system consisting of atoms in the 3D Cartesian space $[\boldsymbol{r}, \boldsymbol{v}]$. The latter is a discrete system expressed as $\boldsymbol{L} = [L_1, L_2]$. The intermolecular distances λ_1 and λ_2 represent reaction coordinates. They are associated, respectively, to the first and second axes of the virtual system: L_1 and L_2. Assume that the system currently takes the virtual state marked by the bold line (green). The other colored states are explained in the legends of following panels. (B) The zone partitioning in the 2D reaction-coordinate space. Each of λ_1 and λ_2 axis is divided into some regions or zones. In this example, five zones are defined for both the axes: $z_1^{(1)}$ through $z_5^{(1)}$ and $z_1^{(2)}$ through $z_5^{(2)}$, where superscript denotes the index for the reaction coordinate (i.e., the dimension of the virtual system), and subscript denotes the index for the zones in each axis. By their combination, there are 25 zones in the 2D $\boldsymbol{\lambda}$ space. (C) When the real system takes λ_1 and λ_2 values at the point "x" in the figure, the virtual system can take one of four virtual states shown as green, pink, cyan, and yellow.

2.3 Construction of the virtual system

The virtual system is defined arbitrarily by users with the following procedure. To begin with, N_{Rc} reaction coordinates $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \cdots, \lambda_{N_{Rc}}]$ are defined as functions of $\boldsymbol{r}$: $\lambda_\gamma = \lambda_\gamma(\boldsymbol{r})$. The γ th reaction coordinate (λ_γ) associates to the γ-th axis of the N_{Rc}-dimensional virtual coordinate $(\boldsymbol{L}_\gamma)$. Then, for each reaction-coordinate axis, the minimum and maximum values defining the range of the reaction coordinate to be sampled are ascertained. This range is divided into subregions or zones. The $\boldsymbol{L}_\gamma$-th zone in the γ-th axis, designated as $\boldsymbol{z}_{L_\gamma}^{(\gamma)}$, is defined as the range of λ_γ, the upper and lower bounds of which are denoted, respectively, as $\left[z_{L_\gamma}^{(\gamma)}\right]_{min}$ and $\left[z_{L_\gamma}^{(\gamma)}\right]_{max}$,

$$\boldsymbol{z}_{L_\gamma}^{(\gamma)} = \left[\left[\boldsymbol{z}_{L_\gamma}^{(\gamma)}\right]_{\boldsymbol{min}}, \left[\boldsymbol{z}_{L_\gamma}^{(\gamma)}\right]_{\boldsymbol{max}}\right]. \tag{2.4}$$

It is noteworthy that neighboring zones should be set to have an intersection. As explained later, the intersection plays a fundamentally important role for migration of the virtual coordinate (i.e., transition among virtual states). Particularly, I applied the condition $\left[z_{L_\gamma}^{(\gamma)}\right]_{max} = \left[z_{L_\gamma+2}^{(\gamma)}\right]_{min}$ to maximize intersections of neighboring zones (Fig. 2-1B). A subspace, or zone, in the N_{Rc}-dimensional reaction-coordinate space associated with the state $\boldsymbol{L}$ is represented as

$$\boldsymbol{z}_{\boldsymbol{L}} = \left[\boldsymbol{z}_{L_1}^{(1)}, \boldsymbol{z}_{L_2}^{(2)}, \dots, \boldsymbol{z}_{L_{N_{Rc}}}^{(N_{Rc})}\right]. \tag{2.5}$$

2.4 Hamiltonian of mD-VcMD

The Hamiltonian of the entire system is defined as the following equation:

$$\mathcal{H} = E_{entire}(\boldsymbol{r}, \boldsymbol{L}) + K(\boldsymbol{v}). \tag{2.6}$$

Kinetic energy $K(\boldsymbol{v})$ does not rely on the virtual system because the virtual particle discretely migrates without the kinetic energy. The potential energy $E_{entire}(\boldsymbol{r}, \boldsymbol{L})$ is described by the following equation:

$$E_{entire}(\boldsymbol{r}, \boldsymbol{L}) = E_R(\boldsymbol{r}) + E_{RV}(\boldsymbol{r}, \boldsymbol{L}) + E_V(\boldsymbol{L}). \tag{2.7}$$

Therein, $E_R(\boldsymbol{r})$ represents the potential energy of the real system defined by the force field. $E_V(\boldsymbol{L})$ is the potential energy of the virtual system for which the currently taken virtual coordinate is $\boldsymbol{L}$. In mD-VcMD method, the potential energy for each virtual state is given as

$$E_V(\boldsymbol{L}) = g_{\boldsymbol{L}}, \tag{2.8}$$

where $\boldsymbol{L}$ denotes a virtual state (the coordinate in the virtual system) and $g_{\boldsymbol{L}}$ is a constant potential energy parameter for the state $\boldsymbol{L}$. Also, $E_{RV}(\boldsymbol{r}, \boldsymbol{L})$ is a coupling term between the real and virtual systems given as a flat-bottom potential:

$$E_{RV}(\boldsymbol{r}, \boldsymbol{L}) = c_{RV} \sum_{\gamma}^{N_{Rc}} \left(\delta_{\lambda_\gamma}\right)^2, \tag{2.9}$$

where

$$\delta_{\lambda_\gamma} = \begin{cases} 0 & (\text{for } \lambda_\gamma \in \boldsymbol{z}_{L_\gamma}^{\gamma}) \\ \lambda_\gamma - \left[z_{L_\gamma}^{(\gamma)}\right]_{min} & \left(\text{for } \lambda_\gamma < \left[z_{L_\gamma}^{(\gamma)}\right]_{min}\right) \\ \lambda_\gamma - \left[z_{L_\gamma}^{(\gamma)}\right]_{max} & \left(\text{for } \lambda_\gamma > \left[z_{L_\gamma}^{(\gamma)}\right]_{max}\right) \end{cases} \tag{2.10}$$

and c_{RV} is a spring constant.

This term restrains the real system in a certain region of the reaction-coordinate space, or zone, associated to the currently taken virtual state: when the system takes a virtual state $\boldsymbol{L}$, motions in the real system are restricted in zone $\boldsymbol{z}_{\boldsymbol{L}}$.

2.5 Migration in each system

In each time step of an mD-VcMD simulation, the system migrates in the phase space via motions of two types: the motion in the real system (CFM) performed by an MD integrator, and motion in the virtual states (inter virtual state transition [IVT]) done by an MC procedure. Later, I provide transition probabilities for the IVT.

For the CFM, the atomic forces are calculated using the derivative of eq. (2.7) with respect to the atomic coordinates as,

$$\text{Forces} = -\frac{\partial E_{entire}(\boldsymbol{r}, \boldsymbol{L})}{\partial \boldsymbol{r}}$$

$$= -\frac{\partial E_R(\boldsymbol{r})}{\partial \boldsymbol{r}} - \frac{\partial E_{RV}(\boldsymbol{r}, \boldsymbol{L})}{\partial \boldsymbol{r}} - \frac{\partial E_V(\boldsymbol{L})}{\partial \boldsymbol{r}}. \tag{2.11}$$

The third term is always zero because it is independent of the atomic coordinates. When reaction coordinate λ is in the zone associated with the current virtual state $\boldsymbol{L}$, the second term is also zero. The system behaves in the same manner as the canonical MD. Otherwise, the second term serves to pull the system to the inside of zone $\boldsymbol{z_L}$.

An IVT is performed by an MC procedure. For a pair of virtual states $\boldsymbol{L}_i$ and $\boldsymbol{L}_j$ with respective zones $\boldsymbol{z}_{\boldsymbol{L}_i}$ and $\boldsymbol{z}_{\boldsymbol{L}_j}$, the IVT between them can occur only when the following condition holds:

$$\boldsymbol{\lambda} \in \left(\boldsymbol{z}_{\boldsymbol{L}_i} \cup \boldsymbol{z}_{\boldsymbol{L}_j}\right). \tag{2.12}$$

Therefore, if $\boldsymbol{\lambda}$ is in the intersection between zones $\boldsymbol{z}_{\boldsymbol{L}_i}$ and $\boldsymbol{z}_{\boldsymbol{L}_j}$, then $\boldsymbol{L}_i$ might transition to $\boldsymbol{L}_j$, assuming that the current virtual state is $\boldsymbol{L}_i$. For the IVT from $\boldsymbol{L}_i$ to $\boldsymbol{L}_j$, the energy difference between before and after IVT is

$$\begin{aligned} \Delta E_{entire}\left(\boldsymbol{L}_j, \boldsymbol{L}_i\right) &= E_V\left(\boldsymbol{L}_j\right) - E_V\left(\boldsymbol{L}_i\right) \\ &= g_{\boldsymbol{L}_j} - g_{\boldsymbol{L}_i}. \end{aligned} \tag{2.13}$$

The energy terms $E_R(\boldsymbol{r})$ and $E_{RV}(\boldsymbol{r}, \boldsymbol{L})$ are zero because an IVT does not change $\boldsymbol{r}$ and does not occur for nonzero $E_{RV}(\boldsymbol{r}, \boldsymbol{L})$ because of eqs. (2.9) and (2.12). Then, the ratio of transition probabilities between these two virtual states is

$$\frac{A_{\boldsymbol{L}_j;\boldsymbol{L}_i}}{A_{\boldsymbol{L}_i;\boldsymbol{L}_j}} = -\exp\left(\frac{\Delta E_{entire}\left(\boldsymbol{L}_j, \boldsymbol{L}_i\right)}{RT}\right), \tag{2.14}$$

where $A_{\boldsymbol{L}_j;\boldsymbol{L}_i}$ denotes the transition probability from $\boldsymbol{L}_i$ to $\boldsymbol{L}_j$. If there are some zones associated with the same intersection (one zone is the current virtual state $\boldsymbol{L}_i$; Fig. 2-1c), then a destination state is chosen randomly according to the transition probabilities (eq. (2.14)), where I use a random number for the selection.

To enhance the conformational sampling, the transition probability should be controlled appropriately by adjustment of the potential parameter $g_{\boldsymbol{L}}$ for each virtual state. A procedure to ascertain the optimal $g_{\boldsymbol{L}}$ is presented in the next section.

2.6 IVT probability

The artificial potential $g_{\boldsymbol{L}}$ is introduced to control the distribution of the resultant ensemble of an mD-VcMD simulation regarding virtual state $\boldsymbol{L}$ as

$$Q_{entire}(\boldsymbol{L}) = \int_{\boldsymbol{\lambda} \in \boldsymbol{z_L}} \rho_{entire}(\boldsymbol{\lambda}, \boldsymbol{L})\, d\boldsymbol{\lambda}, \tag{2.15}$$

where $\rho_{entire}(\boldsymbol{\lambda}, \boldsymbol{L})$ denotes the distribution function of the snapshots at $[\boldsymbol{\lambda}, \boldsymbol{L}]$ in a resultant ensemble. I refer to $Q_{entire}(\boldsymbol{L})$ as the virtual state-partitioned probability. This quantity can be calculable by counting the frequency of snapshots at the state $\boldsymbol{L}$ in trajectories without computation of the integral in eq. (2.15). When the artificial potential $g_{\boldsymbol{L}}$ is a constant irrespective of $\boldsymbol{L}$, an mD-VcMD simulation does not enhance the conformational sampling, and a trajectory behaves similarly to a conventional canonical MD simulation except for snapshots out of zones (with a nonzero $E_{RV}(\boldsymbol{r}, \boldsymbol{L})$). I define the distributions obtained from this unbiased simulation as $\rho_{cano}(\boldsymbol{\lambda}, \boldsymbol{L})$ and $Q_{cano}(\boldsymbol{L})$:

$$\rho_{cano}(\boldsymbol{\lambda}, \boldsymbol{L}) = \rho_{entire}(\boldsymbol{\lambda}, \boldsymbol{L}) \qquad (\text{for } g_{\boldsymbol{L}} = \ const.), \tag{2.16}$$

$$Q_{cano}(\boldsymbol{L}) = Q_{entire}(\boldsymbol{L}) \qquad (\text{for } g_{\boldsymbol{L}} = \ const.). \tag{2.17}$$

I refer to $Q_{cano}(\boldsymbol{L})$ as virtual state-partitioned canonical probability. To enhance the conformational changes, the free-energy gaps between the virtual states $\boldsymbol{L}$ in the unbiased ensemble should be filled with artificial potential $g_{\boldsymbol{L}}$. The potential $g_{\boldsymbol{L}}$ satisfying the following equation

$$-RT \ln\left(\frac{Q_{cano}(\boldsymbol{L}_i)}{Q_{cano}(\boldsymbol{L}_j)}\right) = g_{\boldsymbol{L}_j} - g_{\boldsymbol{L}_i} \tag{2.18}$$

yields a resultant ensemble with the uniform distribution regarding the virtual states as

$$Q_{entire}(\boldsymbol{L}_i) = const. \qquad (\text{for all } \ i). \tag{2.19}$$

Consequently, $Q_{cano}(\boldsymbol{L})$ is necessary to adjust parameter $g_{\boldsymbol{L}}$. In fact, I use only $Q_{cano}(\boldsymbol{L})$ instead of $g_{\boldsymbol{L}}$ for the implementation (see the next section for a method to ascertain the IVT probability $A_{\boldsymbol{L}_j;\boldsymbol{L}_i}$ from $Q_{cano}(\boldsymbol{L})$). Relations among these distribution functions are presented in Figure 2-2.

Because $Q_{cano}(\boldsymbol{L})$ is unknown a priori, this is estimated through iterative simulations of mD-VcMD. For the first iteration, the uniform $g_{\boldsymbol{L}}$ irrespective of $\boldsymbol{L}$ is used. Therefore, the first estimation of $Q_{cano}(\boldsymbol{L})$ is equivalent to the resultant ensemble $Q_{entire}(\boldsymbol{L})$ of the first iteration:

$$Q_{cano}^{[1]}(\boldsymbol{L}) = Q_{entire}^{[1]}(\boldsymbol{L}), \tag{2.20}$$

where superscript [1] represents the quantity obtained from the first iteration. The second iteration is performed using $Q_{cano}^{[1]}(\boldsymbol{L})$; estimation of $Q_{cano}(\boldsymbol{L})$ is updated by reweighting the resultant ensemble as

$$Q_{cano}^{[2]}(\boldsymbol{L}) = Q_{cano}^{[1]}(\boldsymbol{L})Q_{entire}^{[2]}(\boldsymbol{L}). \tag{2.21}$$

In general, the $Q_{cano}(\boldsymbol{L})$ estimated by the M-th iteration is

$$Q_{cano}^{[M]}(\boldsymbol{L}) = Q_{cano}^{[M-1]}(\boldsymbol{L})Q_{entire}^{[M]}(\boldsymbol{L}), \tag{2.22}$$

where $Q_{cano}^{[M]}(\boldsymbol{L})$ is $Q_{entire}(\boldsymbol{L})$ computed from the M-th iteration. If condition eq. (2.19) holds, then $Q_{cano}(\boldsymbol{L})$ is converged. The iterations are performed until obtaining a near-uniform distribution of $Q_{entire}(\boldsymbol{L})$; then the production run is performed using an estimated $Q_{cano}(\boldsymbol{L})$.

In practice, when a virtual state $\boldsymbol{L}_i$ is not sampled in the M-th iteration, $Q_{entire}^{[M]}(\boldsymbol{L}_i)$ and $Q_{cano}^{[M]}(\boldsymbol{L}_i)$ are estimated as zero. To avoid this, I apply a pseudo count replacing the zero $Q_{entire}^{[M]}(\boldsymbol{L}_i)$ with the minimum of $\left\{Q_{entire}^{[M]}(\boldsymbol{L})\right\}$ for sampled virtual states.

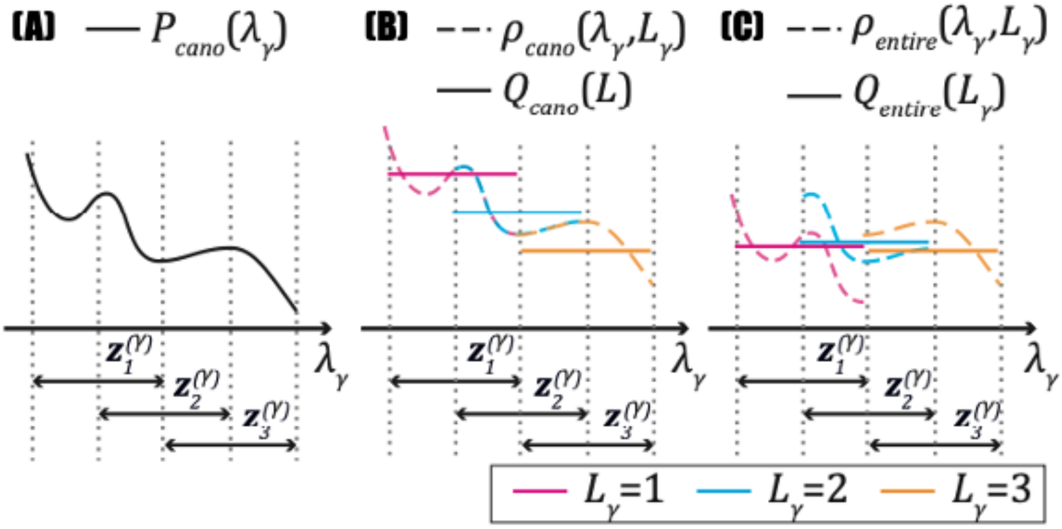

Figure 2-2. Schematic illustrations depict the distributions along a reaction coordinate λ_γ. (A) Canonical distribution $P_{cano}(\lambda_\gamma)$. One purpose of mD-VcMD is estimating this distribution. (B) Distributions obtained from mD-VcMD with constant g_L. The curves of $\rho_{cano}(\lambda_\gamma, L_\gamma)$ for the neighboring L should be well overlapped in a zone intersection, and should be equivalent to $P_{cano}(\lambda_\gamma)$. (C) Distributions obtained from mD-VcMD with biased g_L. Shifting each curve $\rho_{entire}(\lambda_\gamma, L_\gamma)$ by the magnitude of $Q_{cano}(\lambda_\gamma)$ yields $\rho_{cano}(\lambda_\gamma, L_\gamma)$.

2.7 Implementation of IVT

For the implementation of IVT, I do not use g_{L} and eq. (2.14) explicitly. Instead, $Q_{cano}(\boldsymbol{L})$ is used as the parameter. I presume that the system has reaction coordinate $\boldsymbol{\lambda}$, which is at the intersection of n_{link} virtual states, $\boldsymbol{L}^{(1)}, \boldsymbol{L}^{(2)}, \dots, \boldsymbol{L}^{(n_{link})}$. For example, in the 2D reaction-coordinate system, four virtual states can intersect (n_{link} = 4; Fig. 2-1). I refer to set of states having an intersection as linked virtual states. Since the transition probabilities among linked virtual states are independent on properties of virtual states other than these n_{link} virtual states, here I consider only these n_{link} virtual states. I introduce a one-hot vector indicating which virtual state out of n_{link} states is currently taken,

$$\vec{\boldsymbol{\psi}}(\boldsymbol{\lambda}, \boldsymbol{L}) = \begin{bmatrix} \delta\left(\boldsymbol{L} - \boldsymbol{L}^{(1)}\right) \\ \delta\left(\boldsymbol{L} - \boldsymbol{L}^{(2)}\right) \\ \vdots \\ \delta\left(\boldsymbol{L} - \boldsymbol{L}^{(n_{link})}\right) \end{bmatrix}, \tag{2.23}$$

where $\delta\left(\boldsymbol{L} - \boldsymbol{L}^{(i)}\right)$ represents a delta function defined as $\delta\left(\boldsymbol{L} - \boldsymbol{L}^{(i)}\right) = 1$ for the condition $\boldsymbol{L} = \boldsymbol{L}^{(i)}$ and $\delta\left(\boldsymbol{L} - \boldsymbol{L}^{(i)}\right) = 0$ for otherwise. An IVT from $\boldsymbol{\psi}$ to $\boldsymbol{\psi}'$ in each time step is formulated as,

$$\boldsymbol{\psi}' = \boldsymbol{A}\boldsymbol{\psi}, \tag{2.24}$$

where $\boldsymbol{A}$ is a transition probability matrix as

$$\boldsymbol{A} = \begin{bmatrix} A_{L^{(1)};\,L^{(1)}} & A_{L^{(1)};\,L^{(2)}} & \cdots & A_{L^{(1)};\,L^{(n_{link}-1)}} & A_{L^{(1)};\,L^{(n_{link})}} \\ A_{L^{(2)};\,L^{(1)}} & A_{L^{(2)};\,L^{(2)}} & \cdots & A_{L^{(2)};\,L^{(n_{link}-1)}} & A_{L^{(2)};\,L^{(n_{link})}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ A_{L^{(n_{link}-1)};\,L^{(1)}} & A_{L^{(n_{link}-1)};\,L^{(2)}} & \cdots & A_{L^{(n_{link}-1)};\,L^{(n_{link}-1)}} & A_{L^{(n_{link}-1)};\,L^{(n_{link})}} \\ A_{L^{(n_{link})};\,L^{(1)}} & A_{L^{(n_{link})};\,L^{(2)}} & \cdots & A_{L^{(n_{link})};\,L^{(n_{link}-1)}} & A_{L^{(n_{link})};\,L^{(n_{link})}} \end{bmatrix}, \tag{2.25}$$

where probability $A_{L^{(j)};\,L^{(i)}}$, which presents the transition probability from $\boldsymbol{L}^{(i)}$ to $\boldsymbol{L}^{(j)}$, is normalized for each column as $\sum_{j}^{n_{link}} A_{L^{(j)};\,L^{(i)}} = 1$. To perform a VcMD simulation, the probabilities must be specified in advance. To make the resultant virtual state-partitioned probability, $Q_{cano}(\boldsymbol{L_i})$ uniform (eq. (2.19)), the transition probability matrix $\boldsymbol{A}$ is defined as

$$\boldsymbol{A} = \begin{bmatrix} J_{L^{(1)}} & J_{L^{(1)}} & \cdots & J_{L^{(1)}} & J_{L^{(1)}} \\ J_{L^{(2)}} & J_{L^{(2)}} & \cdots & J_{L^{(2)}} & J_{L^{(2)}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ J_{L^{(n_{link}-1)}} & J_{L^{(n_{link}-1)}} & \cdots & J_{L^{(n_{link}-1)}} & J_{L^{(n_{link}-1)}} \\ J_{L^{(n_{link})}} & J_{L^{(n_{link})}} & \cdots & J_{L^{(n_{link})}} & J_{L^{(n_{link})}} \end{bmatrix}, \tag{2.26}$$

where

$$J_{L^{(i)}} = \left[Q_{cano}\left(\boldsymbol{L}^{(i)}\right) \sum_{j=1}^{n_{link}} \frac{1}{Q_{cano}\left(\boldsymbol{L}^{(j)}\right)} \right]^{-1}. \tag{2.27}$$

Consequently, $Q_{cano}(\boldsymbol{L})$ is used as a parameter instead of $g_{\boldsymbol{L}}$ to define the stability of the virtual state $\boldsymbol{L}$.

2.8 Canonical distribution of the real system

A production run of mD-VcMD yields the biased distribution $\rho_{entire}(\boldsymbol{\lambda}, \boldsymbol{L})$. Reweighting this distribution can generate the unbiased canonical distribution in the real system: $\rho_{cano}(\boldsymbol{\lambda})$. To begin with, the weight of each snapshot biased with the artificial potential $g_{\boldsymbol{L}}$ is canceled by multiplying the factor $Q_{cano}(\boldsymbol{L})$ as

$$\rho_{cano}(\boldsymbol{\lambda}, \boldsymbol{L}) = cQ_{cano}(\boldsymbol{L})\rho_{entire}(\boldsymbol{\lambda}, \boldsymbol{L}) \tag{2.28}$$

where c is a normalization factor. For each $\boldsymbol{\lambda}$ in the intersection of n_{link} zones, the weight is averaged over virtual states associated with these linked zones, as

$$P_{cano}(\boldsymbol{\lambda}) = \frac{1}{n_{link}} \sum_{i=1}^{n_{link}} \rho_{cano}\left(\boldsymbol{\lambda}, \boldsymbol{L}^{[i]}\right). \tag{2.29}$$

The canonical distribution $P_{cano}(\boldsymbol{\lambda})$ is derived from the ensemble yielded by an mD-VcMD simulation. It is noteworthy that the canonical ensemble is obtainable based on the weighting parameter $Q_{cano}(\boldsymbol{L})$ irrespective of the convergence of the distribution.

Chapter III

Multidimensional virtual-system coupled canonical molecular dynamics (mD-VcMD) to compute free-energy landscapes of multiple peptides assembly

3.1 Introduction

In this chapter, I provide mD-VcMD's evaluations using simple molecular models consisting of three and four alanine-peptides in explicitly solvated systems. Although these simple systems are not expected to attract biological interest, they are suitable to evaluate the new method. In fact, the conformational ensembles of these systems can be also calculated using the conventional canonical MD method. Therefore, comparison of resultant conformational ensembles between mD-VcMD and the conventional canonical MD can be performed. I demonstrate that these two ensembles including diverse multimer states of peptides are quantitatively consistent. The mD-VcMD method produces the correct canonical ensemble. I adopt intermolecular distances for the multiple reaction coordinates to enhance association and dissociation processes of multimer. I demonstrate a free-energy landscape in the multidimensional reaction coordinate space and conformational diversity analyzed using hierarchical clustering.

3.2 Materials and Methods

3.2.1 Peptide system for 2D-VcMD and 3D-VcMD

As described in "3.1 Introduction" section, I chose simple molecular systems to examine mD-VcMD by direct comparison between quantities from mD-VcMD and those from long-term conventional canonical MD simulations. Because biologically relevant molecular systems are highly complicated and because conformational sampling of such systems is beyond the sampling ability of the conventional canonical MD, simple model systems are necessary for comparison with the conventional method.

I calculated the free-energy landscapes of peptide assemblies for an alanine capped with acetyl (Ace) and N-methyl (Nme) groups. The systems consisting of three and four peptides were simulated. Namely, the system with three capped-alanine (3-

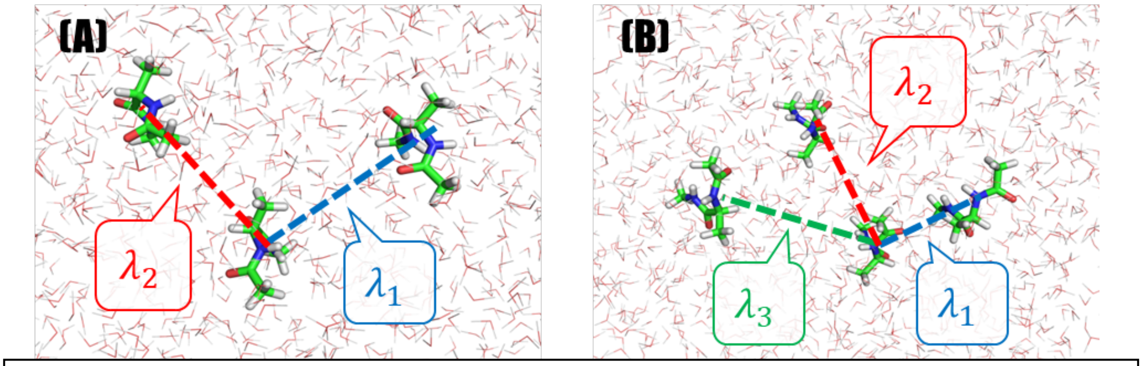

Figure 3-1. Initial conformations of simulations for (A) 3-ALA and (B) 4-ALA systems. Dotted lines represent the reaction coordinates λ_1, λ_2, and λ_3.

ALA), that with four capped-alanine (4-ALA), were simulated. More than one reaction coordinate are necessary to sample various molecular aggregates effectively in a simulation. The actual definition of the reaction coordinates applied here is given later.

To construct the 3-ALA model, a random conformation of an alanine peptide was put in space. Then two copies of the peptide were placed at 10 Å, respectively further along the x-axis and along y-axis. The 4-ALA model was constructed by adding a copy of alanine peptide to the 3-ALA model, where the fourth peptide was at 10 Å distant from the first along the z-axis. Next, the peptides were immersed in a periodic boundary box (40 × 40 × 40 Å^3 for both systems) filled by water molecules with *gmx solvate* module in the GROMACS package. The numbers of atoms were 6,207 and 6,217, respectively, for the 3-ALA and 4-ALA systems.

For each system, energy minimizations with the steepest descent and the conjugate gradient methods were applied. Then, the systems were relaxed through the 1.0 ns MD simulations with gradual heating from 10 to 300 K and an NPT simulation (constant pressure at 1 atm and temperature at 300 K) with the Berendsen barostat. Heavy atoms of the alanine peptides were restrained around the first generated positions during this relaxation process. The resultant cell dimension of the cubic cells were 39.73 Å and 39.83 Å, respectively, for the 3-ALA and 4-ALA systems. Conformations in Figs. 3-1A and 3-1B are the respective final snapshots from the NPT simulations for the 3-ALA and 4-ALA systems.

Their conformational ensembles at 300 K were investigated using the 2D-VcMD and 3D-VcMD simulations. For the 3-ALA system, the first reaction coordinate λ_1 was set to the intermass center distance between the first and second alanine peptides. The second reaction coordinate λ_2 was that between the first and third peptides (Fig. 3-1A). In the case of 4-ALA system, λ_1 and λ_2 were the same as those in the systems with three peptides, and the third reaction coordinate λ_3 was that between the first and fourth peptides (Fig. 3-1B). The 2D and 3D free-energy landscapes are visualized based on the potential of mean force (PMF) along these reaction coordinates. It is noteworthy that mD-VcMD provides statistical weights of each snapshot in the canonical distribution. Therefore, the free-energy landscape can be visualized along arbitrary structural parameters other than the reaction coordinates defining the virtual system. mD-VcMD simulations can be parallelized by using the trivial trajectory parallelization scheme[44]. In this study, 90 independent runs were performed in parallel for each system using different sets of initial atomic velocities. The last snapshot of each of the 90 runs from the M-th iteration was used for the initial conformation for the successive run in the $(M + 1)$-th iteration, except for the first iteration. The first iteration was initiated from the single conformation presented in Figures 3-1A or 3-1B for all runs.

Table 3-1 presents the actual zones for the 2D-VcMD and 3DVcMD simulations. The sampling range of each reaction coordinate was divided into 13 zones and 6 zones, respectively, for 2D-VcMD and 3D-VcMD. I arbitrarily adjusted these conditions to be a similar volume of the virtual systems: 13 × 13 = 169 states and 6 × 6 × 6 = 216 states, respectively, for 2D- and 3D-VcMD simulations. The numbers of iterations were 8 and 7, respectively, for 2D-VcMD and for 3D-VcMD, of which the last iterations were the production runs. The MD time step for the simulation was 2.0 fs. The simulation length is listed in Table 3-2. A snapshot is saved every 5,000 steps of simulation. Consequently, an ensemble of 90,000 snapshots was generated from the production stage for both the 2D-VcMD and 3D-VcMD.

The mD-VcMD simulations were performed using a computer program named myPresto/omegagene[45]. The simulation conditions were the following: the SHAKE algorithm[46] was used to fix the covalent-bond length related to hydrogen atoms, the velocity scaling method[47] was used to control temperature T, and the zero-dipole summation method[48–50] was used to compute electrostatic interactions. The potential energy for the peptides and water molecules were, respectively, the AMBER ff99SB force field[51] and the TIP3P model[52].

Table 3-1. Zone partitioning of each axis

	2D-VcMD		3D-VcMD	
i[a]	$z_{\gamma,i,min}$(Å)	$z_{\gamma,i,max}$(Å)	$z_{\gamma,i,min}$(Å)	$z_{\gamma,i,max}$(Å)
1	3.0	4.0	3.0	5.0
2	3.5	4.5	4.0	6.0
3	4.0	5.0	5.0	7.0
4	4.5	5.5	6.0	8.0
5	5.0	6.0	7.0	9.0
6	5.5	6.5	8.0	15.0[b]
7	6.0	7.0		
8	6.5	7.5		
9	7.0	8.0		
10	7.5	8.5		
11	8.0	9.0		
12	8.5	9.5		
13	9.0	15.0		

[a] Zone index.

[b] For the third axis λ_3, this value is 16.0 Å.

Table 3-2. Simulation length of a single run for each iteration

M	Simulation length (ns)[a]			
	2D-VcMD (3-ALA)	3D-VcMD (4-ALA)	$\mathrm{MD_{conv}}$ (3-ALA)	$\mathrm{MD_{conv}}$ (4-ALA)
1	0.2	0.2	226.2	213.2
2	0.2	0.2		
3	1.0	1.0		
4	1.0	1.0		
5	1.0	1.0		
6	1.0	3.0		
7	3.0	10.0[b]		
8	10.0[b]			

[a] Total simulation length is obtainable by multiplying 90 by the length.
[b] Production stage.

3.2.2 Conventional canonical MD

My earlier study demonstrated that the ensemble obtained using the 1D-VcMD method converges to that from a long-term conventional canonical MD[43].

To confirm the convergence of resultant ensembles obtained from 2D-VcMD and 3D-VcMD to canonical ensemble, I studied the same molecular systems with long-term conventional canonical MD simulations. First, I performed 90 runs of conventional constant volume and temperature MD for all the four systems at 300 K. I designate the conventional MD simulations as $\mathrm{MD_{conv}}$. The simulation lengths of each run are presented in Table 3-2, which are 13 times longer than 2D-VcMD and 3D-VcMD. That is, the total simulation lengths of 90 runs are 20.358 μs and 19.188 μs, respectively, for $\mathrm{MD_{conv}}$ with three peptides (3-ALA) and $\mathrm{MD_{conv}}$ with four peptides (4-ALA), which are 22.6 and 21.3 times longer, respectively, than the production runs of 2D-VcMD and 3D-VcMD simulations, respectively. The flat-bottom potential, eq. (2.9), was used to confine the conformation in the region along all reaction coordinate axes, which is the same region as the 2D-VcMD and 3D-VcMD simulations. The initial conformations of $\mathrm{MD_{conv}}$ were the same as mD-VcMD simulations (Fig. 3-1 for 3-ALA and 4-ALA).

To compare with the results of mD-VcMD, I calculated the virtual state-partitioned canonical probability from $\mathrm{MD_{conv}}$, which is designated as $Q_{MD_{conv}}(\boldsymbol{L})$. This quantity corresponds to $Q_{cano}(\boldsymbol{L})$ from mD-VcMD. Although the conventional MD has no virtual state $\boldsymbol{L}$, $Q_{MD_{conv}}(\boldsymbol{L})$ was calculated as a population of snapshots, $\boldsymbol{\lambda}$ of which is in $\boldsymbol{z_L}$. When a snapshot at an intersection of n_{link} zones, population of all the n_{link} zones are added. Therefore, $\sum_{\boldsymbol{L}} Q_{MD_{conv}}(\boldsymbol{L})$ and $\sum_{\boldsymbol{L}} Q_{cano}(\boldsymbol{L})$ are greater than unity.

3.2.3 Cluster analysis

Conformational diversity of the canonical ensemble obtained from the 3D-VcMD simulation with the 4-ALA system was characterized using a cluster analysis. First, 1,000 representative snapshots were picked from the original ensemble consisting of 90,000 snapshots. These 1,000 snapshots were selected to obey the canonical ensemble $P_{cano}(\boldsymbol{\lambda})$. Next, the topology of intermolecular contacts in each snapshot s was analyzed. To detect contacts, I considered three sites: Ace, Ala, and Nme. Contacts among sites in I-th and J-th peptides were detected based on the threshold r_{sh} for the distance between the centers of mass of the sites and encoded into a 3 × 3 binary matrix $\boldsymbol{C}_{[I;J]}(s)$, with elements $c_{[i,I;j,J]}(s)$. In this study, I applied 7 and 8 Å for r_{sh}. Since a pair of peptides is commutable, this matrix is converted into upper-triangular matrix $\widehat{\boldsymbol{C}}_{[I;J]}(s)$ by adding element $\{i,j\}$ to the element $\{j,i\}$ for $i > j$ pairs, where i and j, respectively, denote indices of row and column in the matrix. In the matrix $\widehat{\boldsymbol{C}}_{[I;J]}(s)$, elements $\hat{c}_{[i,I;j,J]}(s)$ only for $i \leq j$ can have a nonzero value (1 or 2). Since the 4-ALA system includes six I-J pairs, six $\widehat{\boldsymbol{C}}_{[I;J]}(s)$ matrices are definable in a snapshot s. Then, the dissimilarity between a peptide pair $[I;J]$ in a snapshot s_a and a peptide pair $[K;L]$ in a snapshot s_b was defined as the Euclidean distance between two matrices $\widehat{\boldsymbol{C}}_{[I;J]}(s_a)$ and $\widehat{\boldsymbol{C}}_{[K;L]}(s_b)$. I signified this dissimilarity as $D_d(I,J,s_a;K,L,s_b)$. Dissimilarity between two snapshots s_a and s_b was defined as the summation of $D_d(I,J,s_a;K,L,s_b)$ over all six pairs of peptides. The four peptides in the system are chemically identical and not distinguishable. Therefore, there are $4!$ possible values of this dissimilarity between the snapshots derived from the permutation of the peptides. I defined the minimum one as the dissimilarity between the snapshots, $D(s_a,s_b)$. For all pairs of 1,000 snapshots, $D(s_a,s_b)$ were calculated. Hierarchical clustering was performed using Ward's method. Details of this procedure are described in the next subsection.

3.2.4 Inter-snapshot dissimilarity based on intermolecular contact topology

To characterize the resultant ensemble including diverse molecular aggregates, cluster analysis was applied based on an inter-snapshot distance, which is defined here. In this subsection, I introduce a dissimilarity measure based on intermolecular contact networks, disregarding precise molecular and atomic positions; a smaller dissimilarity between two snapshots indicates that fewer contact rearrangements (adding or deleting contacts) are required to transform one network to the other. Because the systems assessed in this study comprise a multimer of chemically identical molecules, the inter-snapshot dissimilarity is expected to be defined as invariant with respect to molecular permutation. This symmetric property originating from the commutability of molecules makes the definition of inter-snapshot dissimilarity complicated.

This subsection provides a definition of dissimilarity between a pair of snapshots. Assuming that the molecular system to be analyzed consists of n_m chemically identical molecules, then for each molecule, n_a sites for intermolecular contacts are defined arbitrarily for example Cα atom of each residue. First, for the snapshot s, inter-site contacts between the i-th site of the I-th molecule and the j-th site of the J-th molecule, which are signified as $c_{[i,I;j,J]}(s)$, are detected for all possible combinations of i, I, j, and J, as

$$c_{[i,I;j,J]}(s) = \begin{cases} 1 & \left(\text{for } r_{[i,I;j,J]} \leq r_{sh}\right) \\ 0 & \quad\text{(otherwise)} \end{cases}, \tag{3.1}$$

where $r_{[i,I;j,J]}$ is the inter-site distance in 3D space and r_{sh} is the distance threshold. Then, site-contact topology between molecules I and J is expressed as a matrix, named a

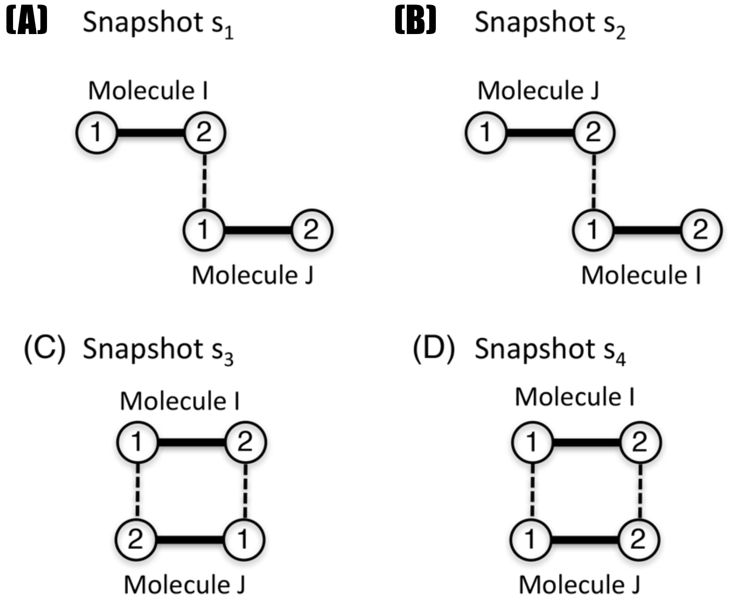

Figure 3-2. Schematic examples of snapshots consisting of two chemically identical molecules indexed as I and J. The four snapshots are termed s_1, s_2, s_3, and s_4, respectively, for panels (A), (B), (C), and (D). Two sites (circles) connected by a solid line constitute a molecule, where digits “1” and “2” are ordinal numbers of the sites. A broken line represents a site contact between two molecules.

site-contact matrix, as

$$\boldsymbol{C}_{[I;J]}(s) = \begin{bmatrix} c_{[1,I;1,J]}(s) & c_{[1,I;2,J]}(s) & \cdots & c_{[1,I;n_a-1,J]}(s) & c_{[1,I;n_a,J]}(s) \\ c_{[2,I;1,J]}(s) & c_{[2,I;2,J]}(s) & \cdots & c_{[2,I;n_a-1,J]}(s) & c_{[2,I;n_a,J]}(s) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{[n_a-1,I;1,J]}(s) & c_{[n_a-1,I;2,J]}(s) & \cdots & c_{[n_a-1,I;n_a-1,J]}(s) & c_{[n_a-1,I;n_a,J]}(s) \\ c_{[n_a,I;1,J]}(s) & c_{[n_a,I;2,J]}(s) & \cdots & c_{[n_a,I;n_a-1,J]}(s) & c_{[n_a,I;n_a,J]}(s) \end{bmatrix} \tag{3.2}$$

where notation $[I;J]$ specifies the pair of molecules and n_a represents the number of sites in a molecule.

Figure 3-2A and 3-2B present the case for a simplified model consisting of two identical molecules (n_m = 2) with two sites for each (n_a = 2). Their site-contact matrices eq. (3.2) are

$$\boldsymbol{C}_{[I;J]}(s_1) = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \text{ and } \boldsymbol{C}_{[I;J]}(s_2) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}. \tag{3.3}$$

However, the intermolecular contact patterns of these two snapshots should be regarded as identical because the molecules are chemically indistinguishable. Therefore, I convert the site-contact matrix $\boldsymbol{C}_{[I;J]}$ to a matrix $\widehat{\boldsymbol{C}}_{[I;J]}$ invariant with respect to the molecular permutation as

$$\widehat{\boldsymbol{C}}_{[I;J]}(s) = \begin{bmatrix} \hat{c}_{[1,I;1,J]}(s) & \hat{c}_{[1,I;2,J]}(s) & \cdots & \hat{c}_{[1,I;n_a-1,J]}(s) & \hat{c}_{[1,I;n_a,J]}(s) \\ 0 & \hat{c}_{[2,I;2,J]}(s) & \cdots & \hat{c}_{[2,I;n_a-1,J]}(s) & \hat{c}_{[2,I;n_a,J]}(s) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \hat{c}_{[n_a-1,I;n_a-1,J]}(s) & \hat{c}_{[n_a-1,I;n_a,J]}(s) \\ 0 & 0 & \cdots & 0 & \hat{c}_{[n_a,I;n_a,J]}(s) \end{bmatrix}, \tag{3.4}$$

where

$$\hat{c}_{[i,I;j,J]}(s) = \begin{cases} c_{[i,I;j,J]}(s) & (\text{for } i = j) \\ c_{[i,I;i,J]}(s) + c_{[j,I;i,J]}(s) & (\text{for } i \neq j) \end{cases}. \tag{3.5}$$

The lower triangular elements of $\widehat{\boldsymbol{C}}_{[I;J]}(s)$, i.e., $\hat{c}_{[i,I;j,J]}(s)$, for $i > j$, are zero. I refer to this matrix as a contact topology matrix. One can define the element for $i \neq j$ as $\left[c_{[i,I;i,J]}(s) + c_{[j,I;i,J]}(s)\right]/2$ in the equation presented above. However, the current definition can express the number of intermolecular contacts clearly as exemplified below (eq. (3.7)). By this definition, both the two site-contact matrices in eq. (3.3) are converted into a contact topology matrix as

$$\widehat{\boldsymbol{C}}_{[I;J]}(s_1) = \widehat{\boldsymbol{C}}_{[I;J]}(s_2) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}. \tag{3.6}$$

The value "1" in the matrix indicates that only one intermolecular contact formed. The site-contact and contact topology matrices for snapshot s_3 in Fig. 3-2C are

$$\boldsymbol{C}_{[I;J]}(s_3) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ and } \widehat{\boldsymbol{C}}_{[I;J]}(s_3) = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix}. \tag{3.7}$$

The value "2" indicates that two intermolecular contacts formed. Fig. 3-2D shows snapshot s_4 as another example, of which the matrices are:

$$\boldsymbol{C}_{[I;J]}(s_4) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } \widehat{\boldsymbol{C}}_{[I;J]}(s_4) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{3.8}$$

Difference between $\widehat{\boldsymbol{C}}_{[I;J]}(s_3)$ and $\widehat{\boldsymbol{C}}_{[I;J]}(s_4)$ is derived from differences in the orientation of molecules. Sites in a molecule are distinguishable when the molecule is a polypeptide. This matrix specifies only whether pairs of sites are in contact or not, irrespective of precise molecular positions. Figure 3-3 presents examples of pairs of snapshots having the same contact topology matrix and different configurations in the 3D space.

Next, I define a dissimilarity measure of the contact topology matrix between a molecular pair $[I;J]$ in snapshot s_a and a pair $[K;L]$ in s_b as the Euclidean distance:

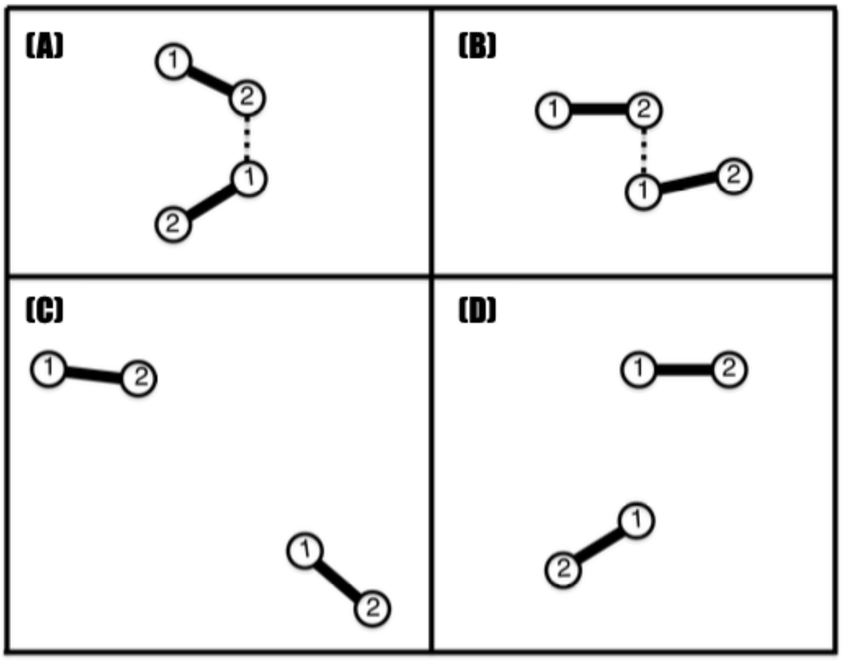

Figure 3-3. Schematic examples of pairs of snapshots with an identical contact topology matrix but different 3D configurations. Whereas panels (A) and (B) have the same contact topology with contact between sites 1 and 2, these snapshots show different orientations of molecules. As another example, panels (C) and (D) with the zero matrix exhibit different relationship of molecules.

$$D_d(I,J,s_a;K,L,s_b) = \left[\sum_{i=1}^{n_a}\sum_{j=1}^{n_a}\left(\hat{c}_{[i,I;\,j,J]}(s_a) - \hat{c}_{[i,K;\,j,L]}(s_b)\right)^2\right]^{1/2} \tag{3.9}$$

Smaller distances represent similar contact topologies between the molecular pairs $[I;J]$ and $[K;L]$. Using eq. (3.9), the dissimilarity among four snapshots s_1, s_2, s_3, and s_4, in Fig. 3-2 is calculated as

$$D_d(I,J,s_1;K,L,s_2) = 0, \tag{3.10}$$

$$D_d(I,J,s_1;K,L,s_3) = D_d(I,J,s_2;K,L,s_3) = 1, \tag{3.11}$$

$$D_d(I,J,s_1;K,L,s_4) = D_d(I,J,s_2;K,L,s_4) = 3^{1/2}, \tag{3.12}$$

$$D_d(I,J,s_3;K,L,s_4) = 6^{1/2}. \tag{3.13}$$

Because snapshots s_1 and s_2 are equivalent, the dissimilarity is zero (eq. (3.10)). In the conformational change from snapshot s_1 to s_3, one contact is added, which results in the dissimilarity of 1 (eq. (3.11)). In the conformational change from s_1 to s_4, two contacts are added with removing one contact. These contact rearrangements correspond to dissimilarity of $3^{1/2}$ (eq. (3.12)). In the conformational change from s_3 to s_4, two contacts are deleted and two contacts are added, for which the dissimilarity is $6^{1/2}$ (eq. (3.13)). The higher dissimilarity indicates that more contact rearrangements are necessary to convert a matrix to the other.

If the system consists of two molecules, then I use eq. (3.9) for the dissimilarity calculation. Next, I consider a system consisting of more than 2 molecules ($n_m > 2$). Figures 3-4A and 3-4B respectively portray two snapshots s_5 and s_6 consisting of four molecules. If the molecules are distinguishable, then the dissimilarity between s_5 and s_6 might be calculated using a summation of $D_d(I,J,s_5;K,L,s_6)$ over the six pairs of $[I;J]$ as

$$\begin{aligned}
D(s_5,s_6) &= \sum_{[I;J]} D_d(I,J,s_5;K,L,s_6) \\
&= D_d(1,2,s_5;1,2,s_6) + D_d(1,3,s_5;1,3,s_6) + D_d(1,4,s_5;1,4,s_6) \\
&\quad + D_d(2,3,s_5;2,3,s_6) + D_d(2,4,s_5;2,4,s_6) + D_d(3,4,s_5;3,4,s_6) \\
&= 1 + (1+4)^{1/2} + 0 + (1+1)^{1/2} + (1+4)^{1/2} + 1 \\
&= 2 + \sqrt{2} + 2\sqrt{5}.
\end{aligned} \tag{3.14}$$

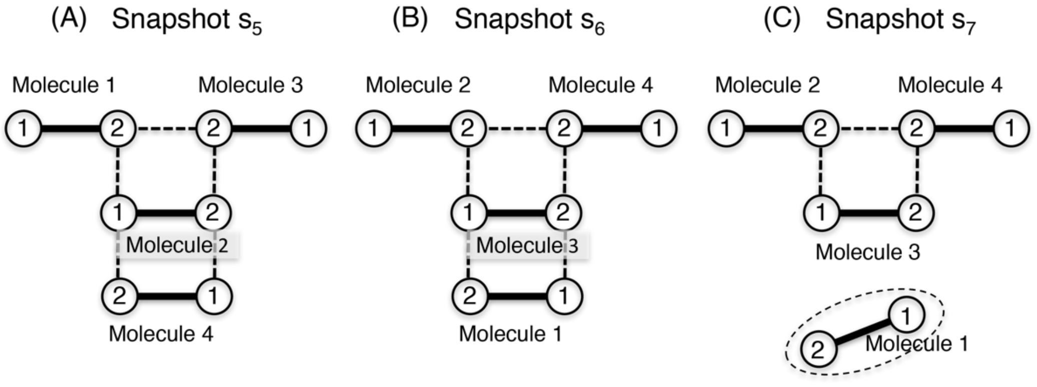

Figure 3-4. Three snapshots s_5 for (A), s_6 for (B), and s_7 for (C). A snapshot consists of four molecules (n_m = 4): molecule 1–4. Broken lines represent contacts between molecules.

In the calculation above, the contact topology matrices for snapshot s_5 (Fig. 3-4A) are

$$\hat{\boldsymbol{C}}_{[1;2]}(s_5) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \hat{\boldsymbol{C}}_{[1;3]}(s_5) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \hat{\boldsymbol{C}}_{[1;4]}(s_5) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\hat{\boldsymbol{C}}_{[2;3]}(s_5) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \hat{\boldsymbol{C}}_{[2;4]}(s_5) = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix}, \hat{\boldsymbol{C}}_{[3;4]}(s_5) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}. \tag{3.15}$$

The contact topology matrices for snapshot s_6 (Fig. 3-4B) are

$$\hat{\boldsymbol{C}}_{[1;2]}(s_6) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \hat{\boldsymbol{C}}_{[1;3]}(s_6) = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix}, \hat{\boldsymbol{C}}_{[1;4]}(s_6) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\hat{\boldsymbol{C}}_{[2;3]}(s_6) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \hat{\boldsymbol{C}}_{[2;4]}(s_6) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \hat{\boldsymbol{C}}_{[3;4]}(s_6) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \tag{3.16}$$

Actually, eq. (3.14) is calculated based on the $[I;J]$-to-$[I;J]$ correspondence between snapshots s_5 and s_6 because it is assumed that all four molecules are mutually distinguished: In other words, a molecular pair $[I;J]$ in snapshot s_5 corresponds to the molecular pair $[I;J]$ in snapshot s_6. For cases in which molecules are indistinguishable, however, many possible correspondences exist between the pairs. For example, a dissimilarity might be defined as

$$D(s_5, s_6) = D_d(1,2,s_5;2,3,s_6) + D_d(1,3,s_5;2,4,s_6) + D_d(1,4,s_5;1,2,s_6) \tag{3.17}$$

$$+D_d(2,3,s_5;3,4,s_6)+D_d(2,4,s_5;1,3,s_6)+D_d(3,4,s_5;1,4,s_6)$$

$$=0+0+0+0+0+0$$

$$=0.$$

This result is natural because the intermolecular contact topology is the same between snapshots s_5 and s_6 in Figs. 3-4A and 3-4B. That is, the two snapshots are equivalent because the molecules are indistinguishable.

Therefore, it is natural to define the inter-snapshot dissimilarity as the minimum of dissimilarities with varying pair correspondence between the snapshots. Given a snapshot, there are $n_m!$ molecular permutations ($n_m! = 24$ for $n_m = 4$), and the k-th arrangement is designated as a_k ($k = 1, \cdots, n_m!$). For instance, the molecular pairs in a_1 and a_2 might be arranged as

$$a_1 = [1;2],[1;3],[1;4],[2;3],[2;4],[3;4], \tag{3.18}$$

and

$$a_2 = [2;3],[2;4],[1;2],[3;4],[1;3],[1;4], \tag{3.19}$$

The arrangement a_1 in eq. (3.18) is one used for snapshots s_6 in eq. (3.17). However, specification of the pair arrangements in a_k is unimportant in these analyses. An important matter is to assign all the $n_m!$ permutations to $\{a_k\}$ exhaustively. I express the six molecular pairs arranged in a_k generally as

$$a_k = a_k(1), a_k(2), a_k(3), a_k(4), a_k(5), a_k(6). \tag{3.20}$$

For the examples in eqs. (3.18) and (3.19), $a_1(3)$ is $[1;4]$, and $a_2(2)$ is $[2;4]$.

Now, I define the inter-snapshot dissimilarity between s_5 arranged in a_1 and s_6 in a_k as

$$D_{[A_1;A_k]}(s_5,s_6) = \sum_{i=1}^{n_m!} D_d\left(a_{1(i)}, s_5; a_{k(i)}, s_6\right)$$

$$= D_d\left(a_{1(1)}, s_5; a_{k(1)}, s_6\right) + D_d\left(a_{1(2)}, s_5; a_{k(2)}, s_6\right) + D_d\left(a_{1(3)}, s_5; a_{k(3)}, s_6\right) \tag{3.21}$$

$$+D_d\left(a_{1(4)}, s_5; a_{k(4)}, s_6\right) + D_d\left(a_{1(5)}, s_5; a_{k(5)}, s_6\right) + D_d\left(a_{1(6)}, s_5; a_{k(6)}, s_6\right),$$

where the $[a_1, a_k]$ specifies that the molecular pairs in snapshots s_5 and s_6 are arranged respectively in a_1 and A_k. Then, the inter-snapshot dissimilarity between snapshots s_5 and s_6 is defined as the minimum of $\{D_{[A_1;\,A_k]}(s_5, s_6)\}$ in all k:

$$D(s_5, s_6) = \min\left[D_{[a_1;\,a_1]}(s_5, s_6), D_{[a_1;\,a_2]}(s_5, s_6), \ldots, D_{[a_1;\,a_{n_m!}]}(s_5, s_6)\right]. \tag{3.22}$$

The general form for the dissimilarity between snapshots s_i and s_j is

$$D(s_i, s_j) = \min_{k=1,\ldots,n_m!}\left[D_{[a_1;\,a_k]}(s_i, s_j)\right]. \tag{3.23}$$

In the equations above, the molecular rearrangement for one snapshot is fixed to a_1 because the other arrangements are redundant for obtaining the minimum.

I consider one more snapshot s_7 (Fig. 3-4C). Molecule 1, surrounded by a broken-line oval, is disconnected from other molecules. Consequently, the dissimilarity between the snapshots s_5 and s_7 is contributed only by the contact deletion as

$$\begin{aligned} D(s_5, s_7) &= \min_{k=1,\ldots,n_m!}\left[D_{[a_1;\,a_k]}(s_5, s_7)\right] \\ &= D_d(2, 4, s_5; 1, 3, s_7) = \sqrt{2}, \end{aligned} \tag{3.24}$$

where

$$\widehat{\boldsymbol{C}}_{[2;4]}(s_5) = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix}, \tag{3.25}$$

and

$$\widehat{\boldsymbol{C}}_{[1;3]}(s_7) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}. \tag{3.26}$$

Once the inter-snapshot dissimilarities are calculated among snapshots in a conformational ensemble, we can generate an *inter-snapshot dissimilarity* matrix $\boldsymbol{D}$ for an ensemble:

$$\boldsymbol{D} = \begin{bmatrix} 0 & D(s_1, s_2) & \cdots & D(s_1, s_{n_s-1}) & D(s_1, s_{n_s}) \\ D(s_2, s_1) & 0 & \cdots & D(s_2, s_{n_s-1}) & D(s_2, s_{n_s}) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ D(s_{n_s-1}, s_1) & D(s_{n_s-1}, s_2) & \cdots & 0 & D(s_{n_s}, s_{n_s}) \\ D(s_{n_s}, s_1) & D(s_{n_s}, s_2) & \cdots & D(s_{n_s}, s_{n_s-1}) & 0 \end{bmatrix} \tag{3.27}$$

Therein, n_s represents the number of snapshots in the ensemble. The diagonal elements are zero: $\Delta(s_i, s_i) = 0$. Applying a clustering method to this inter-snapshot dissimilarity matrix yields clusters based on the intermolecular contact topology are obtained.

3.3 Results and Discussion

3.3.1 Uniformity of raw ensemble of the entire system

The free-energy landscapes of the entire systems sampled from 3-ALA and 4-ALA using the 2D-VcMD and 3D-VcMD methods are shown, respectively, in Figures 3-5A and 3-6A. The landscapes were visualized based on the PMF for each virtual state as

$$PMF_{entire}^{[M]}(\boldsymbol{L}) = -RT\ln\left[Q_{entire}^{[M]}(\boldsymbol{L})\right], \tag{3.28}$$

where superscript $[M]$ represents the M-th iteration. This PMF was normalized so that the largest $Q_{entire}^{[M]}$ is set to 1.0 (the lowest $Q_{entire}^{[M]}$ is set to zero: $PMF_{entire}^{[M]} = 0$). Figures 3-5A and 3-6A, respectively, portray the ensembles obtained from production runs for 3-ALA (2D-VcMD) and 4-ALA (3D-VcMD), that is, eighth and seventh iterations. The landscapes show near-uniform distributions, and eq. (2.19) approximately holds.

To assess the uniformity of distributions, the standard deviation of PMF, $\sigma PMF_{entire}^{[M]}$, for the M-th iteration was evaluated as

$$\sigma PMF_{entire}^{[M]} = \left[\left\langle {PMF_{entire}^{[M]}}^{2}\right\rangle - \left\langle PMF_{entire}^{[M]}\right\rangle^{2}\right]^{1/2}, \tag{3.29}$$

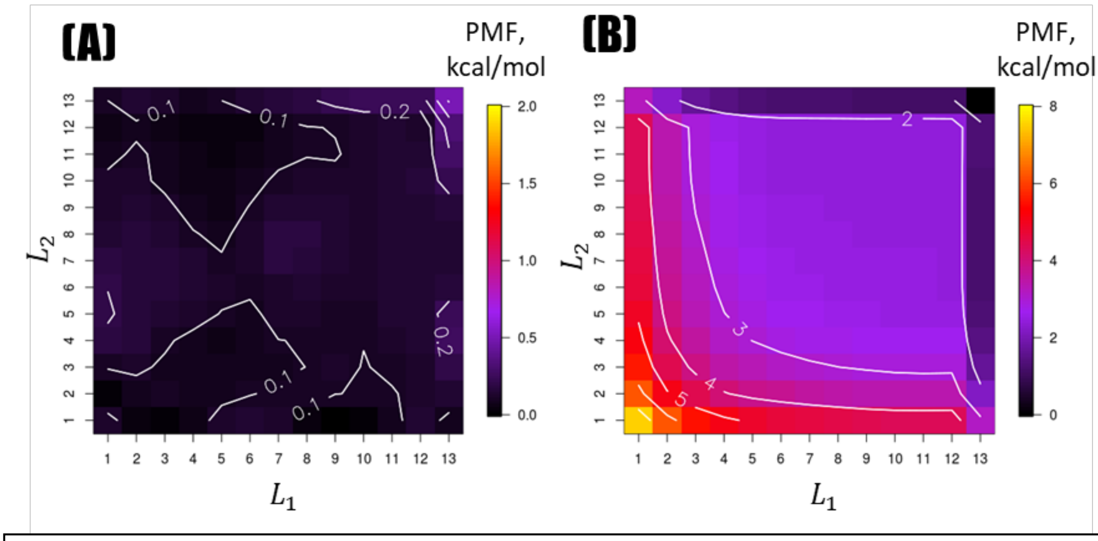

Figure 3-5. Two-dimensional (2D) distributions of the (A) $PMF_{entire}(L_1, L_2)$ and (B) $PMF_{cano}(L_1, L_2)$ for the 3-ALA system calculated from the production run of 2D virtual-system coupled canonical molecular dynamics. The x-axis and y-axis, respectively, represent virtual-state indices L_1 and L_2, which, respectively, correspond to the intermolecular distance between the first and second peptides, and that between the first and third peptides. The PMF values are represented by colors (see the color bars).

where the operator $\langle \cdot \rangle$ denotes the average over the virtual states $\boldsymbol{L}$ sampled in the simulation. Smaller $\sigma PMF_{entire}^{[M]}$ indicates higher uniformity of the virtual state-partitioned probabilities. Figure 3-7A demonstrates that $Q_{entire}^{[M]}(\boldsymbol{L})$ were flattened rapidly with an increase in M for both the 2D-VcMD (3-ALA) and 3DVcMD (4-ALA). The deviations were converged to small values. The $\sigma PMF_{entire}^{[M]}$ values for the production stages of the 3-ALA and 4-ALA systems ($M = 8$ and 7, respectively) were 0.057 and 0.10 kcal/mol, respectively. In theory, because the volume of the conformational space increases drastically with the number of alanine peptides, it is expected that 3D-VcMD provides considerably slower decrease in $\sigma PMF_{entire}^{[M]}$ than 2D-

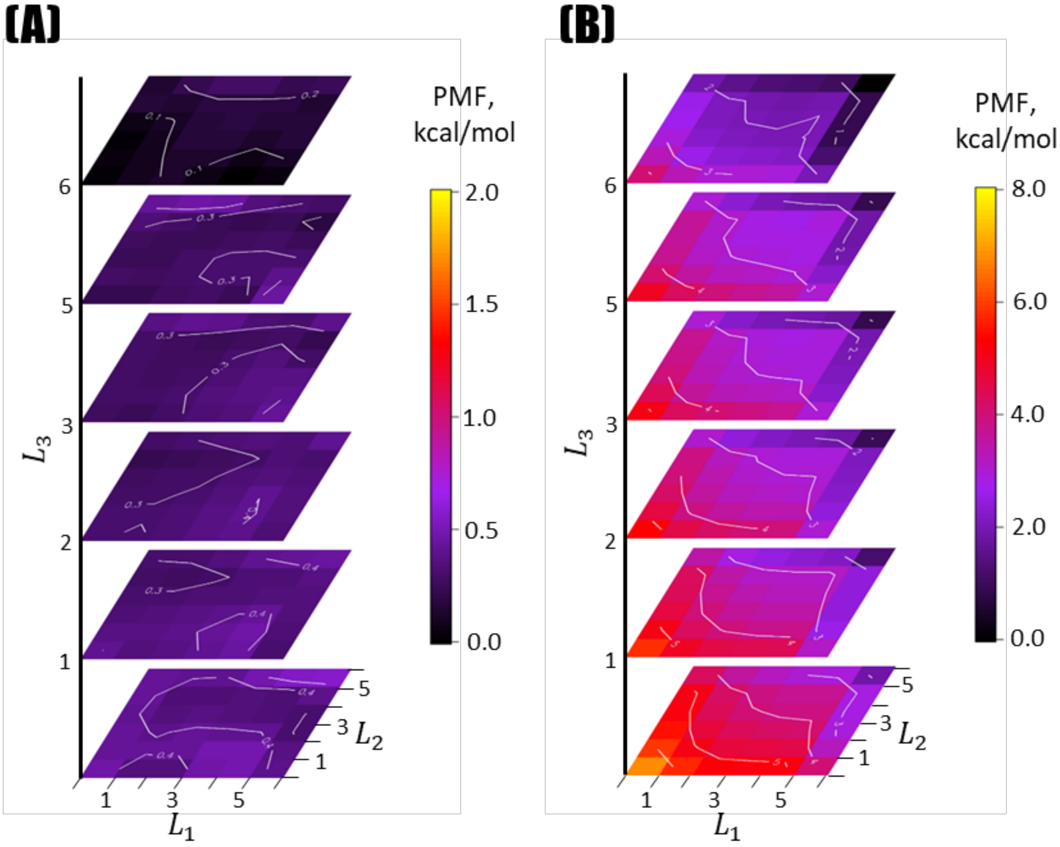

Figure 3-6. Three-dimensional (3D) distributions of the a) $PMF_{entire}(L_1, L_2, L_3)$ and b) $PMF_{cano}(L_1, L_2, L_3)$ for the 4-ALA system calculated from the production run of 3D-virtual-system coupled canonical molecular dynamics. The x-, y-, and z-axes, respectively, represent virtual-state indices L_1, L_2, and L_3, which, respectively, correspond to the intermolecular distance between the first and second peptides, that between the first and third peptides, and that between the first and fourth peptides. The z-axis is shown by slices. The PMF values are shown by colors (see the color bars).

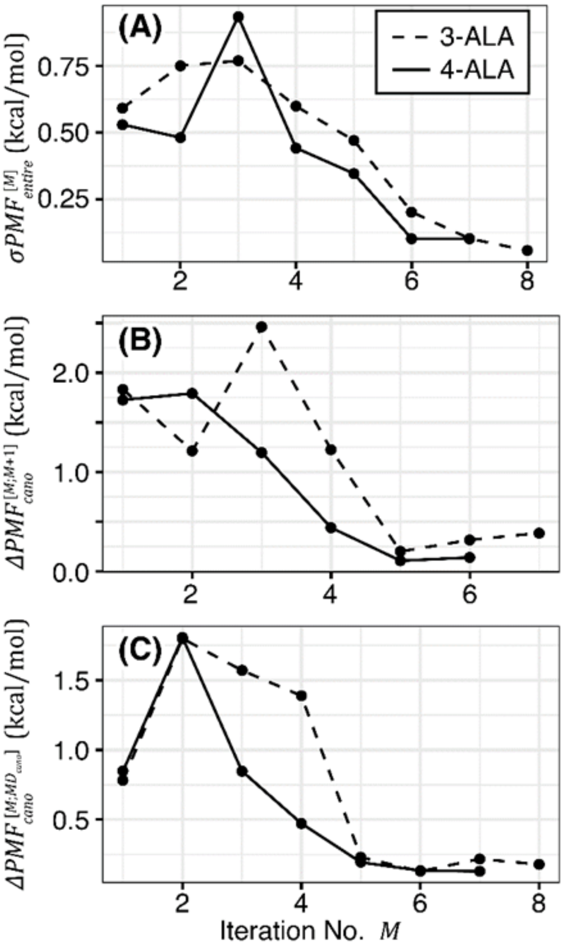

Figure 3-7. Convergence of ensembles along iterations for two-dimensional (2D)-virtual-system coupled canonical molecular dynamics (VcMD; 3-ALA, broken lines) and 3D-VcMD (4-ALA, solid lines). (A) $\Delta PMF_{entire}^{[M]}$ as a function of iteration No. M. (B) $\Delta PMF_{cano}^{[M;M+1]}$ as a function of M. (C) $\Delta PMF_{cano}^{[M;MD_{conv}]}$ as a function of iteration No. M.

VcMD does. However, the convergences were similar between 3D-VcMD and 2D-VcMD (Fig. 3-7A), which demonstrates the effectiveness of the mD-VcMD method to explore a high-dimensional reaction-coordinate space.

3.2.2 Convergence of the canonical ensemble

As described in “3.2 Materials and Methods” section, the distribution of unbiased ensemble $Q_{cano}^{[M]}(\boldsymbol{L})$ is obtainable from the raw distribution $Q_{entire}^{[M]}(\boldsymbol{L})$. The PMF based on the virtual state-partitioned canonical probabilities $Q_{cano}^{[M]}(\boldsymbol{L})$ is defined as

$$PMF_{cano}^{[M]}(\boldsymbol{L}) = -RT \ln\left[Q_{cano}^{[M]}(\boldsymbol{L})\right]. \tag{3.30}$$

This PMF is also normalized so that the largest $Q_{cano}^{[M]}$ has $PMF_{cano}^{[M]} = 0$. Figures 3-5B and 3-6B, respectively, demonstrate the $PMF_{cano}^{[M]}(\boldsymbol{L})$ for the production stages of 2D-VcMD and 3DVcMD. The systems have symmetry for molecular permutations. Therefore, the distribution should be symmetric about the axis permutations in Figures 3-5B and 3-6B. Apparently, the symmetry held well. This result suggests that the sampling was sufficient for both systems. In both the systems, the associative state was less stable than the dissociative state, which occurs because the alanine peptide is too small to form a stable molecular complex without strong charge-charge interactions such as salt-bridges. It is likely that the critical nucleation size for stable growth of a molecular aggregate is larger than four molecules at 300 K at the examined molecular concentration. Convergence of $Q_{cano}^{[M]}(\boldsymbol{L})$ along with iterations was assessed by the root mean square deviation of the PMF between successive iterations:

$$\Delta PMF_{cano}^{[M;M+1]} = \left[\frac{1}{N_{samp}^{[M+1]}} \sum_{\boldsymbol{L}}^{N_{samp}^{[M+1]}} \left[PMF_{cano}^{[M+1]}(\boldsymbol{L}) - PMF_{cano}^{[M]}(\boldsymbol{L}) \right]^2 \right]^{1/2}, \tag{3.31}$$

where $N_{samp}^{[M+1]}$ denotes the number of virtual states sampled in the $(M+1)$-th iteration. If a virtual state $\boldsymbol{L}_i$ was not sampled in the M-th iteration (i.e., $Q_{cano}^{[M]}(\boldsymbol{L}_i) = 0$) and was done in iteration $M+1$ (i.e., $Q_{cano}^{[M+1]}(\boldsymbol{L}_i) > 0$), then I set $Q_{cano}^{[M]}(\boldsymbol{L}_i)$ as the minimum value of $Q_{cano}^{[M]}(\boldsymbol{L})$ in $Q_{cano}^{[M]}(\boldsymbol{L}_i) > 0$. The small $\Delta PMF_{cano}^{[M;M+1]}$ indicates convergence of the virtual state-partitioned probabilities to the stationary state at the M-th iteration.

As a result, Figure 3-7A shows the rapid convergence for both the 3-ALA (2D-VcMD) and 4-ALA (3D-VcMD); deviations between the fifth and sixth iterations were, respectively, $\Delta PMF_{cano}^{[5;\,6]} = 0.20$ kcal/mol and 0.11 kcal/mol. This property of $Q_{cano}^{[M]}(\boldsymbol{L})$ is consistent with the quick convergence of $Q_{entire}^{[M]}(\boldsymbol{L})$ to near-uniform distributions in Figure 3-7A.

3.3.3 Comparison with long-term conventional MD simulation

To confirm the generation of accurate canonical ensembles obtained using the mD-VcMD simulations, I compared the virtual state-partitioned canonical distributions obtained from mD-VcMD, $Q_{cano}^{[M]}(\boldsymbol{L})$, with those obtained from long-term conventional canonical MD simulations, $Q_{MD_{conv}}(\boldsymbol{L})$. It is noteworthy that I have shown convergence of canonical distributions obtained from 1D-VcMD, 1D-VcMC, and 2D-VcMC to those from conventional canonical simulations in earlier studies[24,25,43]. Differences in these distributions were assessed by the root mean square deviations of the PMF by eq. (3.31) with replacing $PMF_{cano}^{[M+1]}(\boldsymbol{L})$ with $PMF_{MD_{conv}}(\boldsymbol{L})$:

$$\Delta PMF_{cano}^{[M;MD_{conv}]} = \left[\frac{1}{N_{samp}^{[M]}} \sum_{\boldsymbol{L}}^{N_{samp}^{[M]}} \left[PMF_{cano}^{[M]}(\boldsymbol{L}) - PMF_{MD_{conv}}(\boldsymbol{L}) \right]^2 \right]^{1/2}, \tag{3.32}$$

where

$$PMF_{MD_{conv}}(\boldsymbol{L}) = -RT\ln\left[Q_{MD_{conv}}(\boldsymbol{L})\right]. \tag{3.33}$$

As a result, $PMF_{cano}^{[M]}(\boldsymbol{L})$ was well converged to $PMF_{MD_{conv}}(\boldsymbol{L})$ in both the systems (Fig. 3-7C). For both the 3-ALA and 4-ALA systems, five iterations were sufficient to converge the distribution; the $\Delta PMF_{cano}^{[M;MD_{conv}]}$ for 3-ALA and 4-ALA at the $M = 5$ were, respectively, 0.23 kcal/mol and 0.19 kcal/mol. Those at the production runs were, respectively, 0.18 kcal/mol and 0.13 kcal/mol, respectively. The cumulative simulation length from the first iteration to the fifth one is 3.4 ns for both the systems (Table 3-2).

In fact, the overall geometry of the free-energy landscape generated by the conventional canonical MD also converged in a similar time scales (Fig. 3-8). However, this result does not necessarily indicate that the conformational sampling of these systems is a trivial issue. The free-energy landscapes (Figs. 3-5B and 3-6B) present that the trimer formation in 3-ALA system and tetramer formation in 4-ALA systems are rare events. In order to assess the efficiency for sampling of such rare states, I analyzed the number of unsampled virtual states in each 1-ns time window of trajectories. Because 90 parallel simulations were performed, a 1-ns window includes 90-ns trajectory. As a result, 81.4% and 82.2% of time windows for 3-ALA and 4-ALA systems had at least one unsampled virtual states in MD_{conv} simulations, those values were zero for 2D- and 3D-VcMD simulations (Fig. 3-9). This clarifies that conformational sampling is highly enhanced by using mD-VcMD method even in such simple molecular systems.

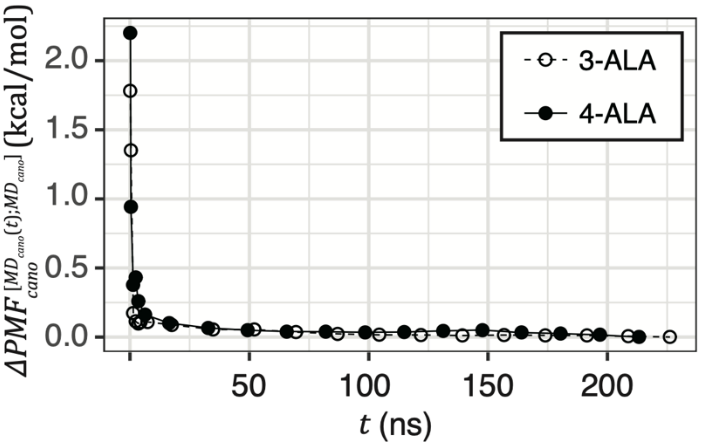

Figure 3-8. Convergence of the ensembles obtained from the MD_{conv} simulations for 3-ALA and 4-ALA systems. The horizontal axis, $\Delta PMF_{cano}^{[MD_{conv}(t);\,MD_{conv}]}$, shows the root mean square deviation of PMF between the ensemble obtained from the trajectories until time t and that from the full-length trajectories.

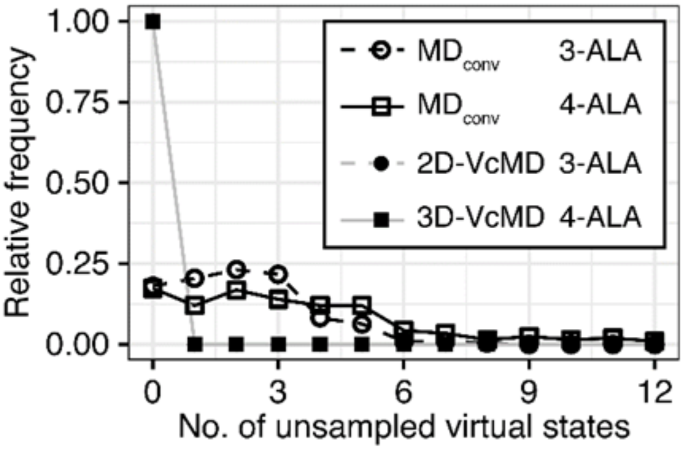

Figure 3-9. Histogram of 1-ns time window over the number of unsampled virtual states out of 169 and 216 states, respectively, for 3-ALA and 4-ALA systems. The results of MD$_{conv}$ (3-ALA), MD$_{conv}$ (4-ALA), two dimensional (2D)-virtual-system coupled canonical molecular dynamics (VcMD; 3-ALA), and 3D-VcMD (4-ALA) are shown in orange, black, red, and blue, respectively. The 2D- and 3D-VcMD simulations resulted in the identical histogram indicating that all the virtual states were sampled in any 1-ns time window.

3.3.4 Conformational diversity in the resultant ensembles

To evaluate details of conformational ensemble generated by 3D-VcMD, I applied a cluster analysis to 1,000 randomly selected snapshots from the resultant ensemble of 4-ALA. Based on the dendrogram depicted in Figure 3-10A, I defined two large clusters termed $c1$ and $c2$. These clusters were divided, respectively, into three and two subclusters, which termed $c1$-1, $c1$-2, $c1$-3, $c2$-1, and $c2$-2 (red frames in Fig. 3-10A; examples of snapshots in each subcluster are shown in Fig. 3-11). These clusters can be characterized roughly in terms of the total number of contacting pairs of peptides, $N_{mol}^{cnt}(s)$, where s denotes the snapshot index. Since the 4-ALA system includes four peptides, $N_{mol}^{cnt}(s)$ takes one integer value from zero to six. Figure 3-10B presents $N_{mol}^{cnt}(s)$ for all representative snapshots. Apparently, most snapshots in cluster $c1$ consisted of large aggregates. Particularly, all the snapshots in the cluster $c1$-3 were of the four-molecule aggregate ($N_{mol}^{cnt}(s) \geq 4$; Fig. 3-11C). Contrarily, most snapshots in $c2$ had fewer than three contacts, which does not form tetrameric aggregates. Particularly, half of the snapshots in $c2$-2 were monomers (the region $N_{mol}^{cnt}(s) = 0$ in Fig. 3-10B; Fig. 3-11E). The intermolecular distances assigned to the reaction coordinates λ_1, λ_2, and λ_3

for each snapshot also presented a trend by which $c1$ forms tighter contacts than $c2$ does, especially for $c1$-3 and $c2$-2 (Fig. 3-10D). Details of contact information in each snapshot are displayed in Figure 3-10C. The vertical axis corresponds to each of 36 positions of intersite contacts, i.e., each element of the upper-triangular matrix $\widehat{\boldsymbol{C}}_{[I;J]}(s)$ for the six pairs of $[I;J]$. The six $\widehat{\boldsymbol{C}}_{[I;J]}(s)$ were arranged in descending order of the total number of contacts (summation of their elements), flattened, and concatenated into a 36-dimensional vector. Figure 3-10C looks sparse even in the snapshot $N_{mol}^{cnt}(s) = 6$ because it is sterically difficult to form tight contacts among all six pairs of peptides.

In summary, the resultant conformational ensemble covered various contacting topologies of the four peptides. This result was robust to choice of the threshold distance to distinguish of whether contact or not; the case r_{sh} = 8.0 Å is shown in Figure 3-12. It is qualitatively consistent with Figure 3-10 with r_{sh} = 7.0 Å.

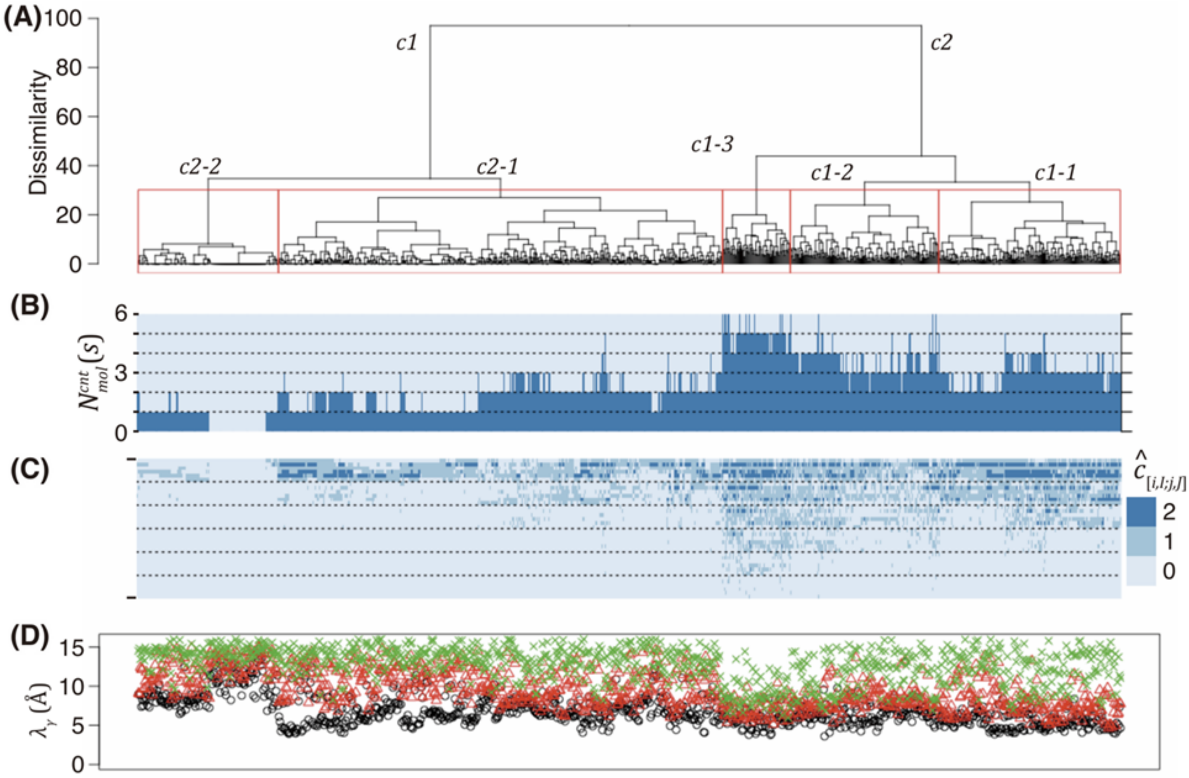

Figure 3-10. Conformational diversity of 1,000 representative snapshots from the 4-ALA system. (A) The dendrogram was calculated based on the contact topology of snapshots. Clusters $c1$ and $c2$ are discriminated at the uppermost stream. Red frames encompass smaller clusters $c1$-1, $c1$-2, …, and $c2$-2 described in the main text. Panels (B), (C), and (D) show characteristics of each snapshot: (B) $N_{mol}^{cnt}(s)$, (C) $c_{[i,I;j,J]}(s)$, which are elements of $\widehat{\boldsymbol{C}}_{[I;J]}(s)$ arranged in the order of the number of contacts in each pair of $[I;J]$ along the y-axis, and (D) the three reaction coordinates, λ_1, λ_2, and λ_3. Black, red, and green circles in panel (D) present the smallest, medium, and highest values of reaction coordinate in the representative snapshots.

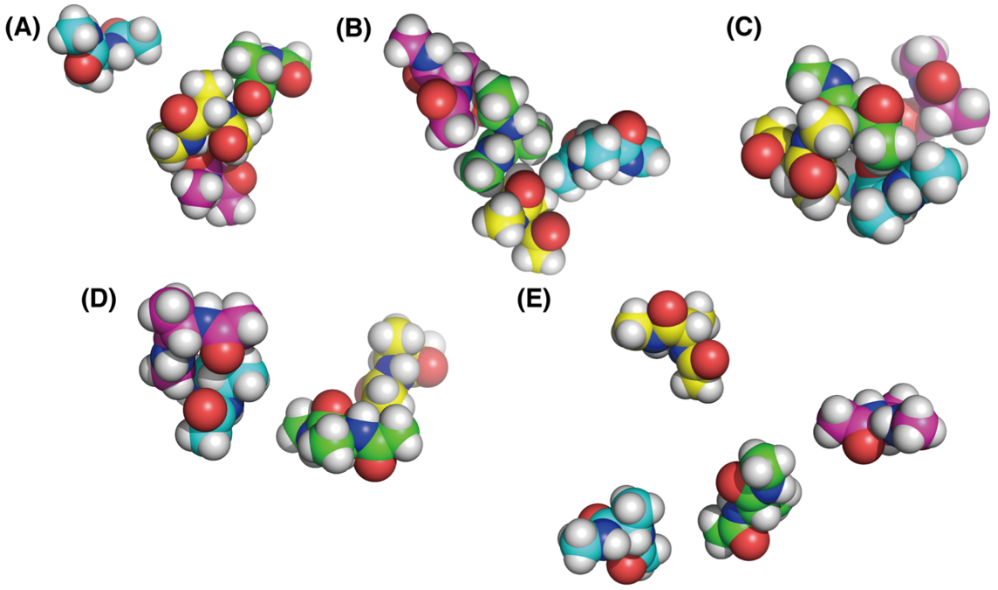

Figure 3-11. Examples of snapshots in the sub-clusters (A) *c*1-1, (B) *c*1-2, (C) *c*1-3, (D) *c*2-1, and (E) *c*2-2c. Carbon atoms in the first, second, third, and fourth alanine peptides are shown, respectively, as green, cyan, magenta, and yellow.

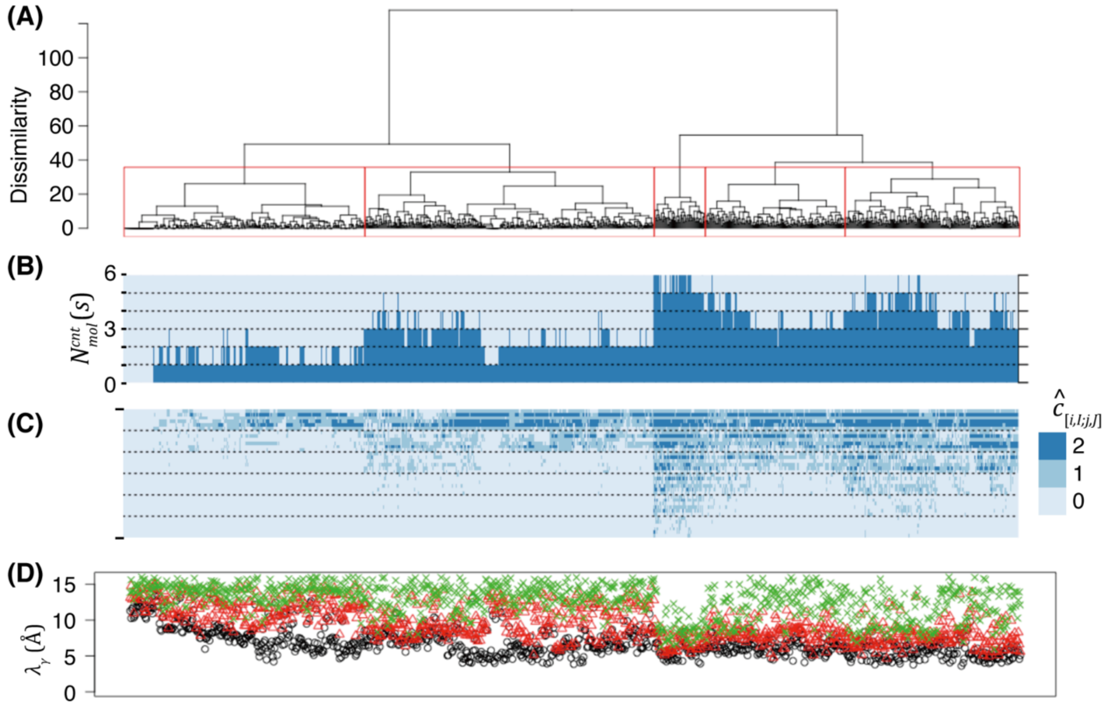

Figure 3-12. Cluster analysis performed with the setting $r_{sh} = 8$ Å for the 4-ALA system. Clustering was applied to the same conformational ensemble used for showing Fig. 3-10 ($r_{sh} = 7$ Å).

3.4 Conclusions

My earlier study of 1D-VcMD demonstrated a conformational sampling enhanced along a single reaction coordinate λ, which is the intermolecular distance between two molecules[43]. For a multimolecular system, however, facilitating rearrangement of multimer formation requires multiple intermolecular distances as reaction coordinates. In this study, I presented the mD-VcMD method to address this issue. Unlike multidimensional AUS, mD-VcMD requires no precise estimation of a high-dimensional canonical distribution function, $P_{cano}(\lambda)$, to calculate the effective potential. Instead, the virtual state-partitioned canonical probability, $Q_{cano}(\boldsymbol{L})$, must be estimated through iterative simulations. $Q_{cano}(\boldsymbol{L})$ can be regarded as a discretized function of $P_{cano}(\lambda)$. The discretization of the distribution function by introducing the virtual system and flat-bottom potentials make it easier to ascertain the effective potential, especially for a multidimensional reaction-coordinate system. In this method, convergence of $Q_{cano}(\boldsymbol{L})$ through iterative simulations realizes uniform sampling in the reaction-coordinate space.

Even if it is not well converged, the canonical ensemble is obtainable by reweighting the raw distribution based on eqs. (2.28) and (2.29), in theory.

Before applying this new method to complicated systems to resolve difficulties related to biology, I examined this method regarding capability to compute accurate canonical ensemble using the simple systems with alanine peptides. The distribution $Q_{cano}(\boldsymbol{L})$ was converged via only a few hundred-scale simulation (3.4 ns for 90 runs for first through fifth iterations in both 2D-VcMD and 3D-VcMD simulations). The resultant ensembles of 2D-VcMD and 3D-VcMD for 3-ALA and 4-ALA systems, respectively, showed quantitative agreement with those provided by long-term (ca. 20 times longer than the VcMD simulations) conventional canonical MD simulations.

Although I presented applications only for simple molecular systems with weak interactions, mD-VcMD is applicable for systems involving stable molecular interaction in which a ligand binds to a deep pocket of a receptor. In such a system, enhancement along the intermolecular distance is insufficient because the ligand cannot enter into or exit from the pocket unless the gate of the pocket loosens. Then, the gate width of the pocket should be taken as another reaction coordinate. This situation was discussed by using a toy model in the earlier study[25]. Analyses of protein–small-ligand binding with a flexible receptor persists as a grand challenge in this field: mD-VcMD is useful to meet this challenge.

For applying complicated systems, to find a suitable definition of a set of reaction coordinates is a key problem for the mD-VcMD method as well as other existing methods, e.g., umbrella sampling. However, as described in "2.1 Introduction" section, discretization of the reaction-coordinate space and biasing with the flat-bottom potential make it easier to apply higher-dimensional reaction-coordinate spaces. Even if it is difficult to find the optimal definition of a single reaction coordinate, applying some multiple reaction coordinates temporarily and finding the optimal coordinates, which can be constructed by linear combinations of the preliminary introduced ones, may provide a practical solution.

Chapter IV

Difference of binding modes among three ligands to a receptor mSin3B corresponding to their inhibitory activities

4.1 Introduction

Neural restrictive silencer factor (NRSF), which is also known as Repressor-element 1 silencing transcription factor (REST)[53,54], is a fundamental repressor, which binds to repressor-element 1 (re1) or neural restrictive silencer element (nrse) of many neuronal genes[55,56]. Importantly, overexpression of NRSF/REST or dysregulation of its cellular expression pattern is related to many neuropathies: Medulloblastoma[57,58], malignant pediatric brain tumor[59], glioblastoma[60,61], Huntington's disease[62–65], neuropathic pain[66,67], Parkinson's disease[68], autism[69], and fibromyalgia[70].

NRSF/REST mediates transcriptional repression recruiting two corepressor complexes: NRSF/REST binds to mSin3 at its N-terminus and to CoREST plus the histone H3K9 methyltransferase G9a at its C-terminus[71]. The mSin3 complex, which contains two histone deacetylases HDAC1 and HDAC2[72], was implicated as an important epigenetic regulator in cancer[73]. The corepressor mSin3B, an isoform of mSin3, consists of four paired amphipathic helix domains (PAH1–PAH4) connected by linkers among the domains, and an intrinsically disordered region of NRSF/REST binds to the cleft of PAH1 of mSin3B[74].

Interestingly, an NMR experiment has shown that the disordered regions of NRSF/REST folds into a helix when binding to the hydrophobic cleft of the PAH1 domain[74] (coupled folding and binding[75–77]). Figure 4-1 shows the PAH1 structure and the cleft position in the domain. A microscopic mechanism for the coupled folding and binding of this system was elucidated by the earlier computational study[18]. The NMR work provided a useful strategy for drug discovery: A compound that inhibits the binding of NRSF/REST to the PAH1 cleft of mSin3B can be a potential drug candidate to ameliorate the neuropathies[70,78–81], and many compounds have been examined[70,74,82].

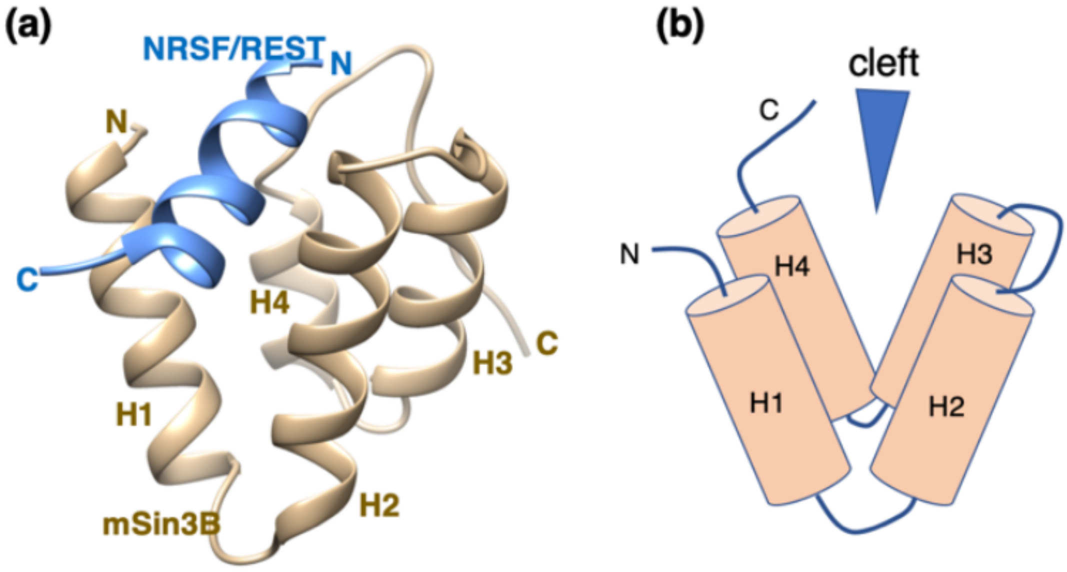

Figure 4-1. (a) Complex structure of PAH1 domain and NRSF/REST. In this chapter, the PAH1 domain of mSin3B is denoted simply as "mSin3B". Four helices, composing mSin3B, are denoted as *H1– H4*. The "C" and "N" are respectively the N- and C-termini for each chain. Shown structure is a snapshot from a sampling simulation[18]. (b) Schematic drawing of the mSin3B structure. The binding cleft of mSin3B is shown by a triangle.

In fact, Ueda et al. analyzed the NMR complex structure of NRSF/REST and the PAH1 domain of mSin3B, and proposed a compound mS-11 to inhibit the binding of NRSF/REST to the PAH1 domain[70]. This compound mimics the helical structure of a four-residue segment (Leu46-Ile47-Met48-Leu49) of NRSF/REST in the bound form, and importantly, this compound inhibited actually the binding of NRSF/REST to the PAH1 domain. I call this four-residue segment a LIML sequence in this study.

As mentioned in Chapters II and III, I developed a generalized ensemble method, multi-dimensional virtual-system coupled molecular dynamics (mD-VcMD) simulation and examined its availability[43,83]. This method enhances conformational sampling of biomolecules in an explicit solvent: By introducing multiple reaction coordinates (RCs) in the molecular system, the conformational motions of the molecules are enhanced with controlling the values of the multiple RCs. The search region in the RC space is expanded through iterative simulation. Importantly, a thermodynamic weight is assigned to each of the sampled conformations, and various thermodynamic quantities of the system are computed from the weighted snapshots. Then, Higo et al. extended the mD-VcMD method by using a genetic algorithm and named the method as a genetic-algorithm-guided mD-VcMD (GA-guided mD-VcMD) simulation[84,85] where the genetic algorithm supports expansion of the search range effectively. They have shown that the sampling efficiency of the GA-guided mD-VcMD is significantly higher than that of the original mD-VcMD[84].

In this chapter, I investigate the spatial distribution of three compounds sertraline, YN3, and acitretin, respectively, around the PAH1 domain of mSin3B, which are obtained from the GA-guided mD-VcMD simulation of the system. The chemical structures of the three compounds are shown in Figure 4-2. I note that preceding experiments on sertraline and YN3 have shown that only sertraline exhibited medulloblastoma cell growth inhibitory activity, although both compounds bound to the PAH1 domain of mSin3B[82]. Another preceding experiment has shown that no inhibitory activities were detected for acitretin whereas it also binds to the PAH1 domain. I show

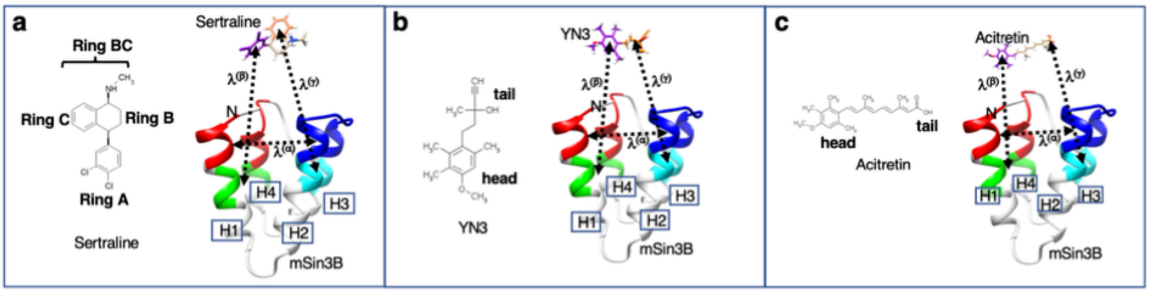

Figure 4-2. (a) (left) Chemical structure of sertraline where three rings are named as "Ring A", "Ring B", "Ring C" in this chapter. Rings B and C are combined and called "Ring BC" simply. (right) Three reaction coordinates (RCs), $\lambda^{(\alpha)}$, $\lambda^{(\beta)}$, and $\lambda^{(\gamma)}$, are introduced in the sertraline–mSin3B system. (b) (left) Chemical structure of YN3. The ring is named as "head", and the opposite side of the ring as "tail" in this chapter. (right) Three reaction coordinates, $\lambda^{(\alpha)}$, $\lambda^{(\beta)}$, and $\lambda^{(\gamma)}$, are introduced in the YN3–mSin3B system. (c) (left) Chemical structure of acitretin. The ring is named as "head", and the opposite side of the ring as "tail" in this chapter. (right) Three reaction coordinates, $\lambda^{(\alpha)}$, $\lambda^{(\beta)}$, and $\lambda^{(\gamma)}$, are introduced in the acitretin–mSin3B system. In each panel, four helices of mSin3B are indicated by "H1", "H2", "H3", and "H4" from the N- to C-terminal. Label "N" shows the position of the N-terminal of mSin3B. $\lambda^{(\alpha)}$ is defined by the distance between the center of mass of red-colored segments of mSin3B and that of blue-colored segment of mSin3B. $\lambda^{(\beta)}$ is defined by the distance between the center of mass of green-colored segments of mSin3B and that of purple-colored part of compound. $\lambda^{(\gamma)}$ is defined by the distance between the center of mass of cyan-colored segments of mSin3B and that of orange-colored part of compound.

that the spatial distribution of the three compounds from the simulations rationally explains why only sertraline exhibited the inhibitory activity. Based on these computational results, I discuss a strategy to develop a drug candidate.

4.2 Materials and Methods

In this chapter, I denote the PAH1 domain of mSin3B simply as "mSin3B" for convenience. Besides, a system composed of sertraline and mSin3B is referred to as a "sertraline–mSin3B" system even when the two molecules are apart to each other during the simulation. Similarly, a system of YN3 and mSin3B is done to as a "YN3–mSin3B" system, and that of acitretin and mSin3B as a "acitretin–mSin3B" system.

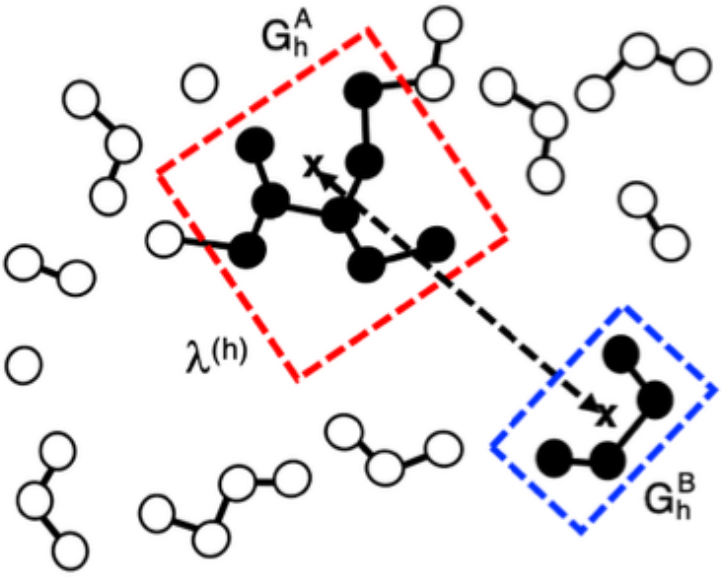

Figure 4-3. Two atom groups, G_h^A and G_h^B, are indicated by, red-colored and blue-colored rectangles, respectively. Atoms in G_h^A and G_h^B are presented by small black filled circles. Center of mass of each atom group is presented by a cross. The distance between the two centers of mass is $\lambda^{(h)}$ (broken-line with arrows).

4.2.1 Three reaction coordinates

First, I introduced multiple RCs in the system, where an RC was defined by the distance between centers of mass of two atom groups. Consider two atom groups G_h^A and G_h^B ($h = \alpha, \beta, \gamma, \ldots$) in a molecular system. The reaction coordinate (RC) $\lambda^{(h)}$ is defined by the distance between centers of mass of G_h^A and G_h^B (Figure 4-3). Superscripts A and B indicate simply that two atom groups are pairing to define $\lambda^{(h)}$, and then, one can exchange the superscripts as: $G_h^A \rightarrow G_h^B$ and $G_h^B \rightarrow G_h^A$ without changing the value of $\lambda^{(h)}$.

For each of the ligand–receptor systems, I introduced three RCs, denoted as $\lambda^{(h)}$ $(h = \alpha, \beta, \gamma)$, presented in Figure 4-2. Schematic representation for the three RCs is given in Figure 4-4a. I briefly explain here the RCs as follows: The two atom groups G_α^A (red-colored segments in Fig. 4-2) and G_α^B (blue-colored segments in Fig. 4-2) define the first RC $\lambda^{(\alpha)}$. We can imagine readily that the move of $\lambda^{(\alpha)}$ opens/closes the cleft. Atom groups G_β^A and G_γ^A are respectively green-colored and cyan-colored segments in mSin3B, and G_β^B and G_γ^B are purple-colored and orange-colored portions in the ligand (Figure 4-2). The moves of $\lambda^{(\beta)}$ (distance between G_β^A and G_β^B) and $\lambda^{(\gamma)}$ (distance between G_γ^A and G_γ^B) control the ligand approaching/departing to mSin3B. When $\lambda^{(\beta)}$ increases with decreasing $\lambda^{(\gamma)}$ or when $\lambda^{(\beta)}$ decreases with increasing $\lambda^{(\gamma)}$, the ligand rotates.

I note that the selection of RCs can be arbitrary in theory if a very long simulation is possible. However, the selection is essentially important to raise the

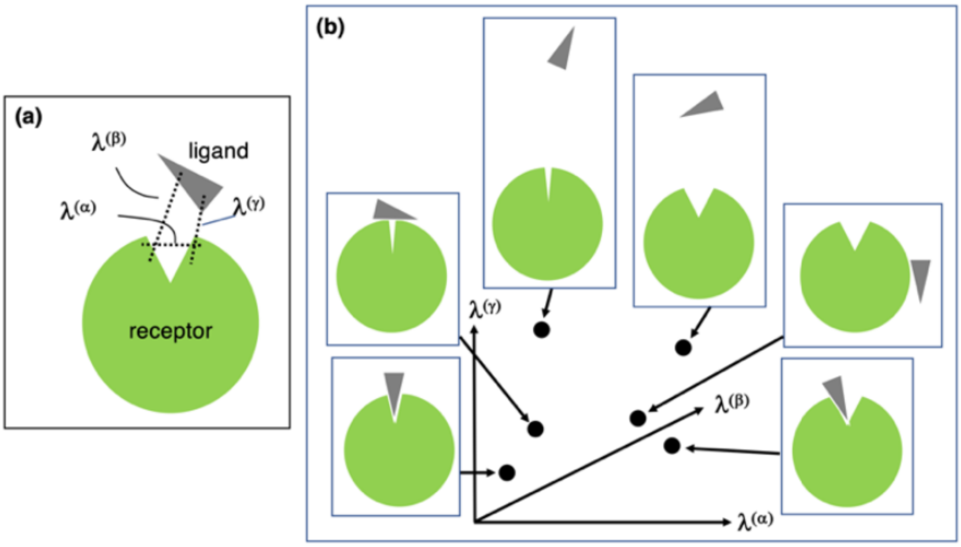

Figure 4-4. (a) Scheme of three RCs, $\lambda^{(h)}$ $(h = \alpha, \beta, \gamma)$, introduced for the current ligand–receptor system: Variation of $\lambda^{(\alpha)}$ controls the cleft opening/closing of the receptor. Variations of $\lambda^{(\beta)}$ and $\lambda^{(\gamma)}$ control ligand approaching/departing from receptor, and relative orientation of the ligand with respect to the receptor. (b) Three-dimensional RC space constructed by $\lambda^{(\alpha)}$, $\lambda^{(\beta)}$, and $\lambda^{(\gamma)}$, where phase point (filled circle) moves according to the system's conformational motion.

efficiency in an actual simulation. Detailed information for the atom groups is given in Table 4-1.

To study molecular binding extensively, the multi-dimensional RC region should involve both the unbound and bound conformations. For this purpose, I set the variable ranges for the three RCs wide enough (Table 4-2). The term “multi-dimensional (mD)” means three-dimensional (3D) in this chapter, whereas the current method is applicable to any dimensional RC space.

Table 4-1. Atom groups to define three RCs.

	Atom Group[a]	
RC	G_h^A	G_h^B
$\lambda^{(\alpha)}$	residues 33–38, 93–99[b] of mSin3B	residues 63–75 of mSin3B
$\lambda^{(\beta)}$	residues 40–43, 90–92 of mSin3B	purple-colored portion in ligand in Fig. 4-2
$\lambda^{(\gamma)}$	residues 60–62, 77–80 of mSin3B	orange-colored portion in ligand in Fig. 4-2

[a] See the main text for definition of atom groups G_h^A and G_h^B $(h = \alpha, \beta, \gamma, \ldots)$. Superscripts A and B are assigned to indicate the atom groups G_h^A and G_h^B, which are pairing to define $\lambda^{(h)}$.

[b] “residues n_1–n_2” involves all atoms in residues n_1–n_2 in mSin3B. See also Fig. 4-2 for positions of the atom groups in mSin3B.

Table 4-2. Parameters[a] for RC-space division

h	$n_{vs}(h)$[a]	$\left[\lambda_1^{(h)}\right]_{min}$[b]	$\left[\lambda_{n_{vs}(h)}^{(h)}\right]_{max}$[b]	$\Delta\lambda^{(h)}$[c]
α	7	10.0 Å	18.0 Å	2.0 Å
β	19	0.0 Å	25.0 Å	2.5 Å
γ	19	0.0 Å	25.0 Å	2.5 Å

[a] Number of virtual states for each reaction coordinate $\lambda^{(h)}$.

[b] Variable range for $\lambda^{(h)}$ is $\left[\left[\lambda_1^{(h)}\right]_{min}, \left[\lambda_{n_{vs}(h)}^{(h)}\right]_{max}\right]$.

[c] Zone width for each reaction coordinate $\lambda^{(h)}$.

4.2.2 Initial conformations of simulation

After setting the RCs above, the initial conformation of simulation was generated. First, the tertiary structure of the receptor mSin3B (PAH1 domain) was taken from the PDBj site (https://pdbj.org/) (PDB ID: 2CZY), in which the receptor binds to the NRSF/REST fragment. After removing NRSF/REST from the complex, I introduced a ligand (sertraline, YN3, or acitretin) near mSin3B. As explained later, I randomized the position of the ligand to generate the initial conformations of the GA-guided mD-VcMD simulation, where the ligand was distant from the binding cleft of mSin3B.

Next, I put the two molecules (ligand and mSin3B) generated above in a periodic boundary box (size is 70.0^3 Å^3) filled by water molecules, and removed water molecules that overlapped to mSin3B or the ligand. Then Na^+ and Cl^- ions were introduced with randomly replacing water molecules by ions. The number of ions was set to a physiological ionic concentration with neutralizing the net charge of the whole system to zero. The resultant sertraline–mSin3B system consists of 33,543 atoms (1,200 atoms for mSin3B, 38 atoms for sertraline, 10,749 water molecules, 29 Na^+, 29 Cl^-), the YN3–mSin3B system does of 33,533 atoms (1,200 atoms for mSin3B, 40 atoms for YN3, 10,745 water molecules, 29 Na^+, 29 Cl^-), and the acitretin–mSin3B system does of 33,525 atoms (1,200 atoms for mSin3B, 50 atoms for acitretin, 10,739 water molecules, 29 Na^+, 29 Cl^-). After a short energy minimization, a short constant-volume and constant-temperature (300 K) simulation (NVT simulation) was performed. Then, a constant-pressure (1 atm) and constant-temperature (300 K) simulation (NPT simulation) was performed to relax the box size. The resultant box size was 68.920^3 Å^3, 68.909^3 Å^3, and 68.909^3 Å^3 for the sertraline–mSin3B, YN3–mSin3B, and acitretin–mSin3B systems, respectively. Those computations were done by using a program package myPresto/psygene[86]. The force fields used for those simulations are described later.

Whereas the PAH1 domain is linked to the PAH2 domain by a long flexible linker in a real cell, only the PAH1 domain was computed in this simulation, and the inter-domain linker was treated as the C-terminal tail. This tail might be inserted into the binding cleft of PAH1 domain incidentally during the simulation. It is likely that the incidental insertion of the inter-domain linker into the cleft does not happen if the PAH1 and PAH domains are connected by the linker. Thus, to prevent this incidental and artificial event, I applied weak restraints to the C-terminal tail. I introduced distance-restraint energy, E_{res}, between a part of PAH1 domain of mSin3B and its C-terminal tail to prevent the C-terminal tail from being inserted into the binding cleft of PAH1 domain. The function form is:

$$E_{res} = \begin{cases} 0.5 \sum_{i,j} \left[r_{i,j} - \left(r_{i,j}^{0} - r_{low} \right) \right]^2 & (\text{for } r_{i,j} \leq r_{i,j}^{0} - r_{low}) \\ 0 & (\text{for } r_{i,j}^{0} - r_{low} < r_{i,j} < r_{i,j}^{0} + r_{up}), \\ 0.5 \sum_{i,j} \left[r_{i,j} - \left(r_{i,j}^{0} + r_{up} \right) \right]^2 & (\text{for } r_{i,j} \geq r_{i,j}^{0} + r_{up}) \end{cases} \tag{4.1}$$

where $r_{i,j}$ and $r_{i,j}^{0}$ are the C$_\alpha$ atomic distance between residues i and j in a simulation snapshot and the reference complex structure (the NMR structure of the NRSF/REST–mSin3B complex; PDB ID: 2CZY), respectively, and r_{low} and r_{up} specify tolerances set to 2.0 Å. Thus, no restraint ($E_{res} = 0$) is applied to the atom-pair distance $r_{i,j}$ when $r_{i,j}^{0} - r_{low} < r_{i,j} < r_{i,j}^{0} + r_{up}$. This distance restraint was applied to three C$_\alpha$ atomic distance pairs between Gly 92 and Asp 104, between Phe 93 and Ile 105, and between Asn 94 and Arg 106. By the restraints, the C-terminal did not move to the binding cleft, and fluctuated around the initial conformation (NMR structure; PDB ID: 2CZY) during the simulation.

To generate the initial conformations of simulation, where the ligand is distant from the binding cleft of mSin3B, I applied interactions between mSin3B and the ligand so that the RCs fall in the following ranges: 15 Å $< \lambda^{(\alpha)} <$ 16 Å, 24 Å $< \lambda^{(\beta)} <$ 25 Å, and 24 Å $< \lambda^{(\gamma)} <$ 25 Å. Then, with applying these interactions I performed 256 runs starting from the last snapshot of the NPT simulation done above. Figure 4-5 display some of the last conformations picked randomly from those 256 runs for the three systems. I used those 256 conformations for the initial conformation of GA-guided mD-VcMD. Apparently, the ligand in these conformations was distant from the cleft of mSin3B (i.e., NRSF/REST position). These figures also display the NRSF/REST fragment binding to the cleft of mSin3B (PDB ID: 2CZY), whereas NRSF/REST did not exist in the current simulation.

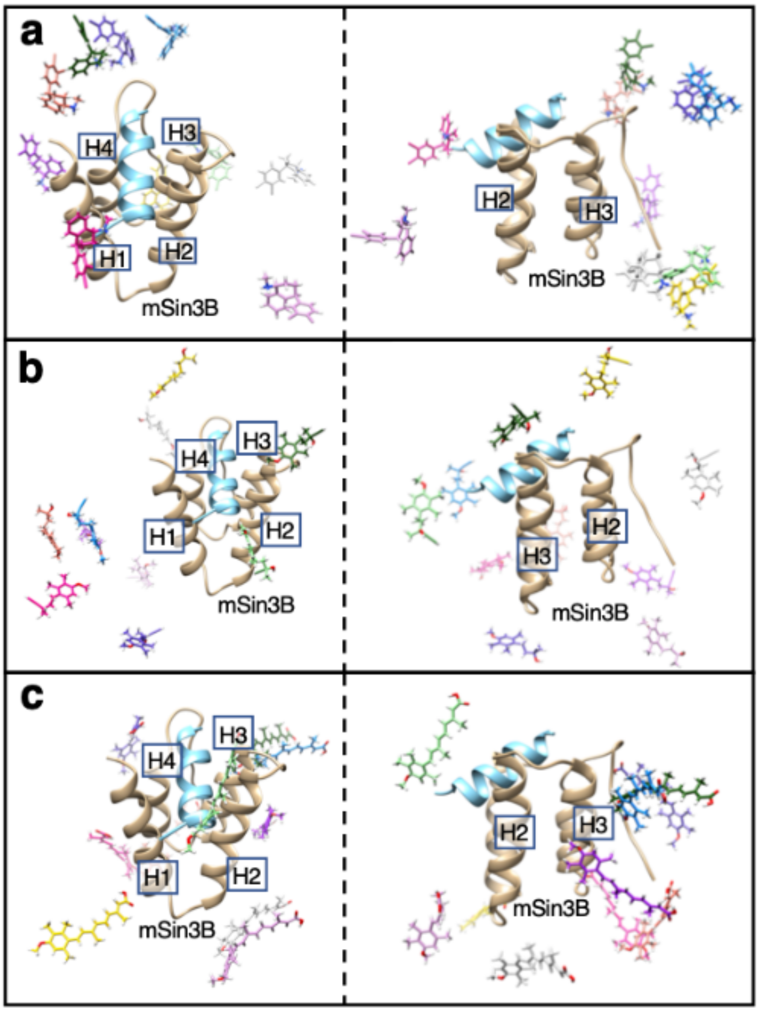

Figure 4-5. (a) Some initial conformations of GA-guided mD-VcMD for the sertraline–mSin3B system viewed from two different directions. Solvent is omitted. Shown mSin3B structure is the NMR structure (PDB ID: 2CZY). This figure also displays the NRSF/REST fragment (cyan-colored model) binding to mSin3B in the NMR structure, while NRSF/REST does not exist in the GA-guided mD-VcMD simulation. Labels H1, ..., H4 are helices 1–4 of mSin3B (PAH1 domain). (b) Some initial conformations for the YN3–mSin3B system viewed from two different directions. (c) Some initial conformations for the acitretin–mSin3B system viewed from different directions.

4.2.3 The GA-guided mD-VcMD

In order to determine IVT probability, $Q_{cano}(\boldsymbol{L})$ is used as the parameter (eqs. (2.26) and (2.27)). $Q_{cano}(\boldsymbol{L})$ is estimated through iterative simulations of mD-VcMD, and $Q_{entire}(\boldsymbol{L})$

is used to update $Q_{cano}(\boldsymbol{L})$ (eqs. (2.21) and (2.22)). $Q_{entire}(\boldsymbol{L})$, the virtual state-partitioned probability, is calculated by counting the frequency of snapshots at the state $\boldsymbol{L}$ in trajectories. Therefore, if a poorly sampled region in the multidimensional RC space arose in an iterative simulation, this region might lead to conformational entrapment in a later iteration. To avoid the difficulty, Higo et. al. proposed a subzone-based mD-VcMD method, which is an extension of the original mD-VcMD; and a GA-guided mD-VcMD method, which extends the subzone-based mD-VcMD to expand sampling to non-sampled RC regions using genetic algorithm (GA)[84].

Methodological details for GA-guided mD-VcMD are explained in Ref 84. Here, I explain the outline of GA-guided mD-VcMD and resultant quantities. This method controls the system's motions by modulating the three RCs $\lambda^{(h)}$ $(h = \alpha, \beta, \gamma)$. Figure 4-4b presents schematically a distribution of the system's conformation in the 3D RC space resulted from the moves of $\lambda^{(h)}$.

The outline of the method is as follows: First, the entire 3D RC space is divided into many small zones. Then, the GA-guided mD-VcMD simulation provides a conformational distribution function $Q_{cano}\left(\lambda^{(\alpha)}, \lambda^{(\beta)}, \lambda^{(\gamma)}\right)$ of the system, where $Q_{cano}\left(\lambda^{(\alpha)}, \lambda^{(\beta)}, \lambda^{(\gamma)}\right)$ is the probability of existence at position $\left[\lambda^{(\alpha)}, \lambda^{(\beta)}, \lambda^{(\gamma)}\right]$ in the 3D RC space constructed by $\lambda^{(\alpha)}$, $\lambda^{(\beta)}$, and $\lambda^{(\gamma)}$. Because the conformational space of a biological molecular system is wide, the GA-guided mD-VcMD is executed via iterative simulations, during which the sampled RC region is expanded. The simulation is terminated when $Q_{cano}\left(\lambda^{(\alpha)}, \lambda^{(\beta)}, \lambda^{(\gamma)}\right)$ has converged. I discuss the convergence quantitatively later. After the convergence, a thermodynamic weight is assigned to each of stored snapshots using $Q_{cano}\left(\lambda^{(\alpha)}, \lambda^{(\beta)}, \lambda^{(\gamma)}\right)$, and the ensemble of the snapshots can be regarded as a thermally equilibrated conformational ensemble (canonical ensemble)[84]. If the GA-guided mD-VcMD simulation is done at temperature T, then the canonical ensemble at T is obtained.

I performed 256 runs for an iteration in the present study to raise the sampling efficiency further. A simple integration of the 256 trajectories can be regarded as a long single trajectory[44,87]. When the M-th iteration is finished, I have snapshots stored from the 1st to M-th iterations. The 256 initial conformations for the $(M + 1)$-th iteration were selected from those stored snapshots so that the conformations distributed as even as possible in the 3D-RC space. On the other hand, when an RC region was sampled poorly, I prepared the initial conformations around the poorly sampled region[84]. Because I obtained the canonical ensemble at 300 K as a result of the GA-guided mD-VcMD, I calculated the distribution function of various quantities in equilibrium at 300 K.

4.2.4 Simulations

The atom groups adopted for the current study to define the three RCs are given in Table 4-1. Three RCs are illustrated in Figure 4-2. GA-guided mD-VcMD consists of iterative simulations, through which the conformational ensemble converges on an equilibrated one, $Q_{cano}(\boldsymbol{\lambda})$ $\left(\boldsymbol{\lambda} = \lambda^{(\alpha)}, \lambda^{(\beta)}, \lambda^{(\gamma)}\right)$, at a simulation temperature (300 K) in the 3D-RC space. A thermodynamic weight is assigned to each of stored snapshots using $Q_{cano}(\boldsymbol{\lambda})$ [84,85]. This means that a thermally equilibrated conformational ensemble (canonical ensemble) is obtained in the allowed RC space. Table 4-2 lists actual values of parameters which control the simulations: $n_{vs}(h)$, $\left[\lambda_1^{(h)}\right]_{min}$, $\left[\lambda_{n_{vs}(h)}^{(h)}\right]_{max}$, and $\Delta\lambda^{(h)}$. Table 4-3 lists the zones: $\left\{\left[\lambda_k^{(h)}\right]_{min}, \left[\lambda_k^{(h)}\right]_{max}; k = 1, \ldots, n_{vs}(h)\right\}$. An interzone transition was attempted once every 20 ps (1×10^4 steps of simulation). See Ref. 84 for meaning of parameters. The simulation was performed using a computer program myPresto/ omegagene [45] with the following condition: SHAKE[46] to fix the covalent-bond lengths related to hydrogen atoms, the Berendsen thermostat to control temperature[47], the zero-dipole summation method[48–50] to compute accurately and quickly the long-range electrostatic interactions, a time-step of 2 fs ($\Delta t = 2$ fs), and simulation temperature of 300 K. The Berendsen thermostat produces an ensemble that can approximate a canonical distribution for a many-atom system, whereas it generates a non-physical distribution for a small system[88]. To compute the original potential energy of the system, the Amber hybrid force fields (mixture parameter $w = 0.75$)[89] was used for mSin3B, the TIP3P potential model for water molecules[52], and the Joung–Cheatham model for chloride and sodium ions[90].

The force fields for the sertraline, YN3, and acitretin were set as follows: First, the atomic partial charges were derived by quantum chemical calculations using Gaussian03[91] at the HF/6-31G* level, followed by RESP fitting[92]. Then, those partial charges were incorporated into a GAFF (general amber force field) force-field file[93]. GAFF was designed to be compatible with conventional AMBER force-fields. The Amber hybrid force fields currently used for mSin3B was generated with mixing Amber parm94[94] and parm96[95] force fields to treat both helical and stranded polypeptides[89], and the difference between parm94 and parm96 exists only in the dihedral energy parameters. Therefore, the inter-molecular interaction energy between mSin3B and the ligands is invariant mechanically among the parm94, parm96, and hybrid force fields. Those force field parameters were used for the energy minimization, NVT, NPT, and the GA-guided mD-VcMD simulations.

Table 4-3. Setting of zones

Zone No.[a]	Zones[b]					
k	$\left[\lambda_k^{(\alpha)}\right]_{min}$	$\left[\lambda_k^{(\alpha)}\right]_{max}$ [c]	$\left[\lambda_k^{(\beta)}\right]_{min}$	$\left[\lambda_k^{(\beta)}\right]_{max}$	$\left[\lambda_k^{(\gamma)}\right]_{min}$	$\left[\lambda_k^{(\gamma)}\right]_{max}$
1	10.0	12.0	0.00	2.50	0.00	2.50
2	11.0	13.0	1.25	3.75	1.25	3.75
3	12.0	14.0	2.50	5.00	2.50	5.00
4	13.0	15.0	3.75	6.25	3.75	6.25
5	14.0	16.0	5.00	7.50	5.00	7.50
6	15.0	17.0	6.25	8.75	6.25	8.75
7	16.0	18.0	7.50	10.00	7.50	10.00
8			8.75	11.25	8.75	11.25
9			10.00	12.50	10.00	12.50
10			11.25	13.75	11.25	13.75
11			12.50	15.00	12.50	15.00
12			13.75	16.25	13.75	16.25
13			15.00	17.50	15.00	17.50
14			16.25	18.75	16.25	18.75
15			17.50	20.00	17.50	20.00
16			18.75	21.25	18.75	21.25
17			20.00	22.50	20.00	22.50
18			21.25	23.75	21.25	23.75
19			22.50	25.00	22.50	25.00

[a] Number of zones n_{vs} is 7 for $\lambda^{(\alpha)}$ and 19 for $\lambda^{(\beta)}$ and $\lambda^{(\gamma)}$.

[b] Unit of zones is Å.

[c] $\left[\lambda_k^{(\alpha)}\right]_{min}$ $\left[\lambda_k^{(\alpha)}\right]_{man}$ are not assigned for $k \geq 8$ because number of virtual states is 7: $n_{vs}(\alpha) = 7$. See Table 4-2.

4.2.5 Spatial density of a compound around mSin3B

As mentioned above, the GA-guided VcMD simulation produces a conformational ensemble, where a thermodynamic weight at 300 K is assigned to each constituent conformation. Thus, I can calculate a spatial distribution function of any structural quantity from the ensemble. In this study, I compute the spatial density $\rho_{CG}^{(s)}(\boldsymbol{r})$, which is the probability of detecting the geometrical center (GC) of the ligand in the vicinity of a three-dimensional position $\boldsymbol{r} = [x, y, z]$ in the real space and the superscript s is the system specifier: $s =$ sertraline–mSin3B, YN3–mSin3B, or acitretin–mSin3B.

The GA-guided mD-VcMD assigns a thermodynamic weight (statistical weight at equilibrium) to each sampled snapshot[84]. Here I present a method to calculate spatial distribution of the center of mass of the ligands around the receptor mSin3B. First, I divide the 3D real space into cubes, whose volume ΔV is 2 Å × 2 Å × 2 Å. The cube position is specified by its center $\boldsymbol{r} = [x, y, z]$. Next, I calculate the geometrical center (GC) of the ligand for each snapshot, and assign the snapshot to a cube that involves the GC. Then, I assign all of the snapshots to cubes. Last, I calculate the spatial density $\rho_{GC}(\boldsymbol{r})$ of the geometrical center at the cube $\boldsymbol{r}$ as:

$$\rho_{CG}^{(s)}(\boldsymbol{r}) = \sum_i w_i \delta_{GC}(\boldsymbol{r}; i), \tag{4.2}$$

where w_i is the thermodynamic weight assigned to snapshot i, and the superscript s specifies the ligand: s = sertraline–mSin3B, YN3–mSin3B, or acitretin–mSin3B. The function $\delta_{GC}(\boldsymbol{r}; i)$ is a delta function defined:

$$\delta_{GC}(\boldsymbol{r}; i) = \begin{cases} 1 & (\text{if CG of snapshot } i \text{ is in cube } \boldsymbol{r}) \\ 0 & (\text{else}) \end{cases}. \tag{4.3}$$

Equation (4.2) is the thermodynamic weight assigned to the cube $\boldsymbol{r}$. I normalized $\{w_i\}$ as $\sum_i w_i = 1$ for each system in advance.

Another spatial density function is calculated with the same manner. For instance, a spatial density $\rho_{GCA}(\boldsymbol{r})$ for the geometrical center of Ring A of sertraline (GCA) is calculated as follows:

$$\rho_{CGA}^{(s)}(\boldsymbol{r}) = \sum_i w_i \delta_{GCA}(\boldsymbol{r}; i), \tag{4.4}$$

where

$$\delta_{GCA}(\boldsymbol{r}; i) = \begin{cases} 1 & (\text{if CGA of snapshot } i \text{ is in cube } \boldsymbol{r}) \\ 0 & (\text{else}) \end{cases}. \tag{4.5}$$

The superscript s is set to sertraline–mSin3B.

4.3 Results

4.3.1 Distribution of the system's conformation in 3D RC space

I repeated 14, 13, and 27 iterations for the sertraline–mSin3B, YN3–mSin3B, and acitretin–mSin3B systems, respectively. For all of the three systems, an iteration was composed of 256 runs. Each run was performed for 3×10^6 steps (6 ns; time step $\Delta t = 2$ fs). Thus, the total simulation length was 21.504 μs (= $14 \times 256 \times 6$ ns), 19.968 μs (= $13 \times 256 \times 6$ ns), and 41.472 μs (= $27 \times 256 \times 6$ ns) for the sertraline–mSin3B, YN3–mSin3B, and acitretin–mSin3B systems, respectively. A snapshot was stored every 1×10^5 steps (200 ps), yielding 107,520, 99,840, and 207,360 snapshots, for the sertraline–mSin3B, YN3-mSin3B, and acitretin–mSin3B systems, respectively.

Figure 4-6 demonstrates the conformational distributions of the three systems in the 3D RC space, where the density is normalized so that the largest density is set to 1.0. I access the convergence of $Q_{cano}\left(\lambda^{(\alpha)}, \lambda^{(\beta)}, \lambda^{(\gamma)}\right)$ with proceeding the iteration (i.e., the ordinal number of iteration). The assessment is done with $E_{local}\left(\lambda^{(\alpha)}, \lambda^{(\beta)}, \lambda^{(\gamma)}\right)^{-1}$ that is a function defined locally at each position in 3D RC space. This function was introduced to assess the accuracy of $Q_{cano}\left(\lambda^{(\alpha)}, \lambda^{(\beta)}, \lambda^{(\gamma)}\right)$[84], and used actually to check the simulation quality[85]. The larger the function $E_{local}{}^{-1}$ at a position $\left[\lambda^{(\alpha)}, \lambda^{(\beta)}, \lambda^{(\gamma)}\right]$, the better the accuracy of Q_{cano} at the position. According to Ref. 85, I judged that a 3D-RC region with $E_{local}{}^{-1} \geq 4.0$ has an appropriate accuracy.

Figures 4-7, 4-8, and 4-9 indicate that the regions with $Q_{cano}\left(\lambda^{(\alpha)}, \lambda^{(\beta)}, \lambda^{(\gamma)}\right) > 0.001$ are converged well. Although Fig. 4-6 is basically important to show that the sampling covered a wide conformational space, this figure is not useful for understanding the ligand's distribution around mSin3B. In the next subsection, I analyze the ligand's distribution using the canonical conformational ensemble.

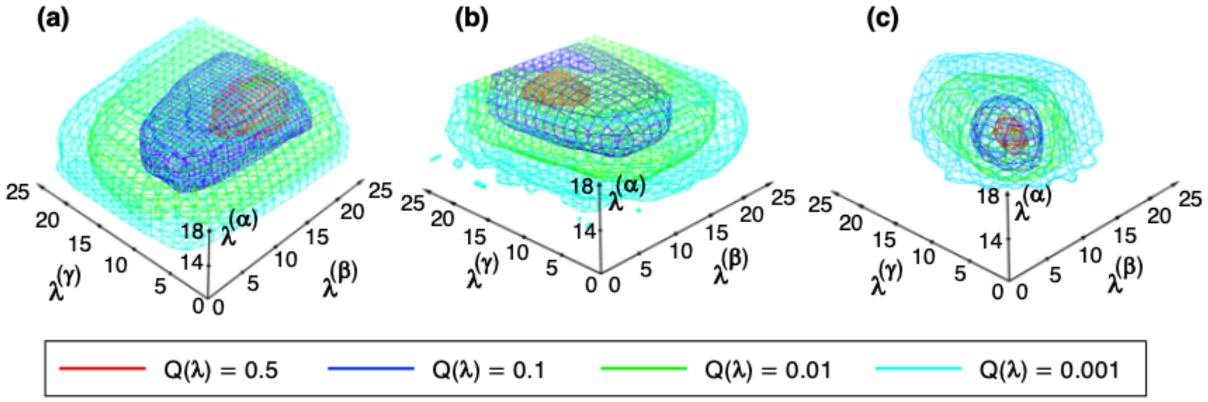

Figure 4-6. Density $Q_{cano}\left(\lambda^{(\alpha)}, \lambda^{(\beta)}, \lambda^{(\gamma)}\right)$ of (a) the sertraline–mSin3B, (b) YN3–mSin3B, and (c) acitretin–mSin3B systems in the 3D-RC space. In GA-guided mD-VcMD, the distribution is defined originally by $Q_{cano}\left(L^{(\alpha)}, L^{(\beta)}, L^{(\gamma)}\right)$, where $L^{(\alpha)}$, $L^{(\beta)}$, and $L^{(\gamma)}$ are respectively indices to specify the positions $\lambda^{(\alpha)}$, $\lambda^{(\beta)}$, and $\lambda^{(\gamma)}$ in the 3D-RC space. Then, I convert $\left[L_i^{(\alpha)}, L_j^{(\beta)}, L_k^{(\gamma)}\right]$ to: $\lambda_i^{(h)} = 0.5\left\{\left[\lambda_i^{(h)}\right]_{min} + \left[\lambda_i^{(h)}\right]_{max}\right\}$, where $i = 1, \ldots, n_{vs}(h)$ $(h = \alpha, \beta, \gamma)$. See Table 4-3 for values of $\left[\lambda_i^{(h)}\right]_{min}$, $\left[\lambda_i^{(h)}\right]_{max}$, and $n_{vs}(h)$. Then, $Q_{cano}\left(\lambda^{(\alpha)}, \lambda^{(\beta)}, \lambda^{(\gamma)}\right)$ is normalized so that the highest density is set to 1. Contour levels are presented by colors in inset.

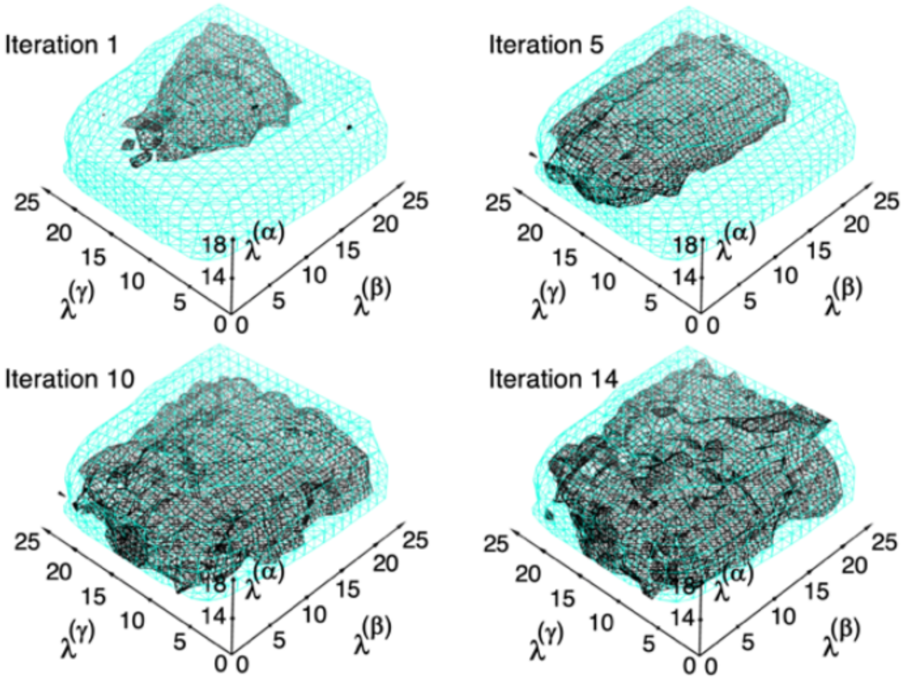

Figure 4-7. Convergence of distribution $Q_{cano}(\lambda^{(\alpha)},\lambda^{(\beta)},\lambda^{(\gamma)})$ with proceeding iterative simulation for the sertraline–mSin3B system. Accuracy of $Q_{cano}(\lambda^{(\alpha)},\lambda^{(\beta)},\lambda^{(\gamma)})$ is assessed by objective function $E_{local}(\lambda^{(\alpha)},\lambda^{(\beta)},\lambda^{(\gamma)})^{-1}$ [84]. Cyan contours, i.e., the equidensity region of $Q_{cano}(\lambda^{(\alpha)},\lambda^{(\beta)},\lambda^{(\gamma)}) = 0.001$, are those used in Figure 4-6. Black contours are the iso-objective-function surfaces of $E_{local}(\lambda^{(\alpha)},\lambda^{(\beta)},\lambda^{(\gamma)})^{-1} = 4$. The iteration No. of each panel is indicated near the panel.

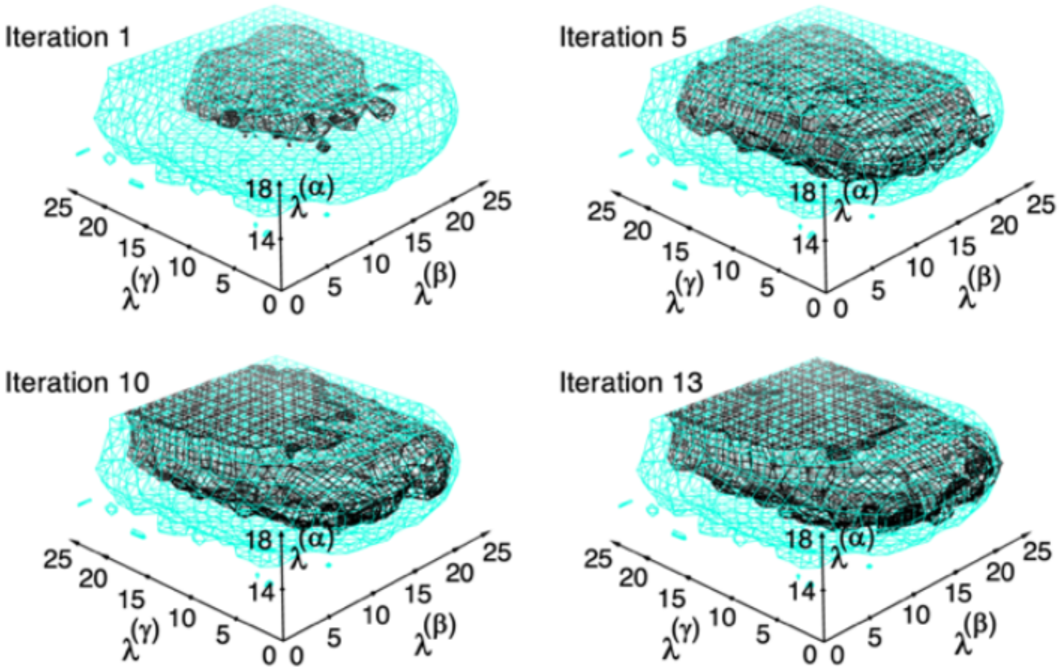

Figure 4-8. Convergence of distribution $Q_{cano}\left(\lambda^{(\alpha)},\lambda^{(\beta)},\lambda^{(\gamma)}\right)$ with proceeding iterative simulation for the YN3–mSin3B system. See caption of Figure 4-7 for more information.

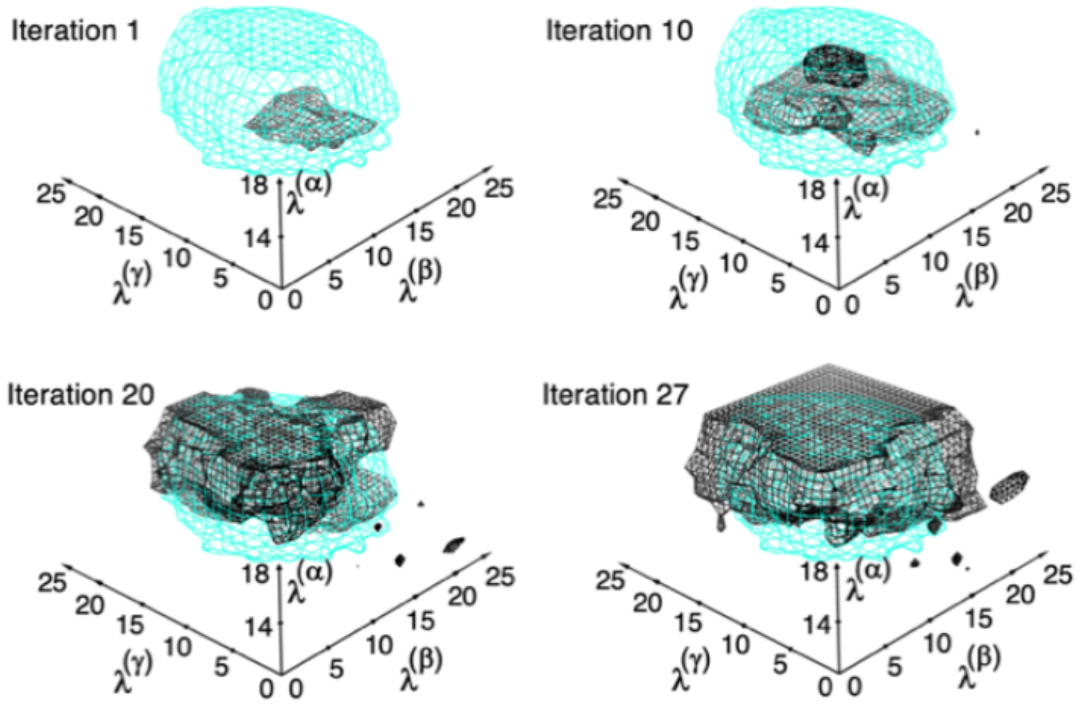

Figure 4-9. Convergence of distribution $Q_{cano}(\lambda^{(\alpha)},\lambda^{(\beta)},\lambda^{(\gamma)})$ with proceeding iterative simulation for the acitretin–mSin3B system. See caption of Figure 4-7 for more information.

4.3.2 Distribution of ligands around mSin3B

I computed the spatial density $\rho_{CG}^{(s)}(\boldsymbol{r})$ for the geometric center of the ligands (Fig. 4-10). At the low-contour level ($\rho_{CG}^{(s)}(\boldsymbol{r}) = 0.005\rho_0$; green-colored contours), the ligands distributed almost everywhere around mSin3B for all systems. This result is natural indicating that the sampling was done widely. With increasing the density level, the ligands tended to be localized at some surface regions of mSin3B ($\rho_{CG}^{(s)}(\boldsymbol{r}) = 0.05\rho_0$; blue-colored contours). At the high-density level ($\rho_{CG}^{(s)}(\boldsymbol{r}) = 0.005\rho_0$; red-colored contours), a high-density conformational cluster (labeled Cluster A in the figure) can be identified in the cleft of mSin3B for sertraline and YN3, whereas acitretin did not show a remarkable cluster in the cleft. This indicates that the ligand–mSin3B binding for sertraline and YN3 is stronger than that for acitretin.

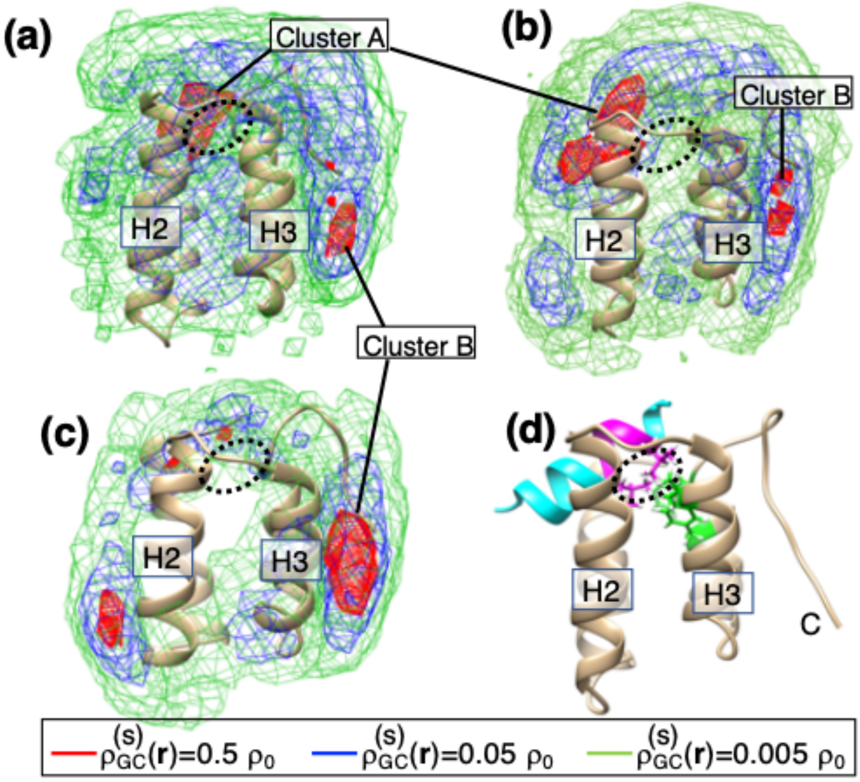
(a)
Cluster A
(b)
Cluster B
H2
H3
Cluster B
(c)
(d)
H2
H3
H2
H3
H2
H3
C
$\rho^{(s)}_{GC}(\mathbf{r})=0.5\ \rho_0$
$\rho^{(s)}_{GC}(\mathbf{r})=0.05\ \rho_0$
$\rho^{(s)}_{GC}(\mathbf{r})=0.005\ \rho_0$

Figure 4-10. Spatial density $\rho_{CG}^{(s)}(\boldsymbol{r})$ (s = sertraline, YN3, or acitretin) of the geometric center (GC) at position $\boldsymbol{r}$ for (a) the sertraline–mSin3B, (b) YN3–mSin3B, and (c) acitretin–mSin3B systems in the 3D real space. See Section 4.2.5 for procedure to calculate $\rho_{CG}^{(s)}(\boldsymbol{r})$. Contour levels are shown in inset where $\rho_0 = 0.001$. Displayed structure of mSin3B is that after NPT simulation for each system, where labels H2 and H3 represent helix 2 and helix 3 of mSi3B, respectively. The high-density cluster ($\rho_{CG}^{(s)}(\boldsymbol{r}) > 0.5\rho_0$) in the cleft of mSin3B (PAH1) is named as Cluster A, and one near the N-terminal of mSin3B as Cluster B. (d) NMR structure of NRSF/REST–mSin3B complex (PDB ID: 2CZY), where cyan-colored model is NRSF/REST, and the magenta-colored segment is the LIML sequence of NRSF/REST. Label C indicates the position of the C-terminal tail of mSin3B. Two magenta-colored sidechains are Leu 46 and Leu 49 of the LIML sequence. Black broken-line circle indicates the position of the two sidechains. The circles in panels (a), (b), and (c) are presented to indicate the sidechain position of Leu 46 and Leu 49. Green-colored sidechains are Val 75, Phe 93, and Phe 96 of mSin3B (see also green-colored sidechains of Figure 4-11a), which form a hydrophobic core with Leu 46 and Leu 49 of the LIML sequence.

Figure 4-10d displays the NRSF/REST–mSin3B complex structure solved by the NMR experiment from the same orientation of Fig. 4-10a–c. A view of the NRSF/REST–mSin3B complex from a different orientation is presented in Figure 4-11. The binding cleft of mSin3B is hydrophobic (Figure 4-11b), and two residues Leu 46 and Leu 49 of the LIML sequence bind to the cleft forming a hydrophobic core with Val 75, Phe 93, and Phe 96 of mSin3B. See "4.1 Introduction" section for the LIML sequence. Interestingly, only the geometrical center of sertraline overlaps the position of Leu 46 and Leu 49 of the LIML sequence (black broken-line circle) in the NRSF/REST–mSin3B complex, when the spatial density is viewed at the contour level of $\rho_{CG}^{(s)}(\boldsymbol{r}) = 0.5\rho_0$. Therefore, sertraline forms a tighter hydrophobic core with the binding cleft of mSin3B than the other ligands do.

I also observed a high-density cluster (Cluster B in Fig. 4-10) near the C-terminal tail of mSin3B. Because the C-terminal tail was restrained (see the "4.2 Materials and Methods"), the C-terminal tail did not fluctuate largely in the simulation in spite that the C-terminal tail is exposed to solvent. Therefore, it is likely that Cluster B was induced by this less-fluctuating C-terminal tail: Cluster B is an artificial cluster. Figure 4-12 displays a sertraline's conformation taken from Cluster B. The nitrogen atom of Ring B of sertraline interacts electrostatically to an oxygen atom of Asp 104 of the C-terminal tail. Ring A of sertraline interacts with the hydrophobic residue Ile 61 of helix H4 of mSin3B, and Ring C of sertraline does with the hydrophobic residue Val 62 of helix H4. Majority of conformations from Cluster B showed those interaction patterns. Therefore, it is likely that Cluster B disappears when the C-terminal tail is highly exposed to solvent and fluctuates largely.

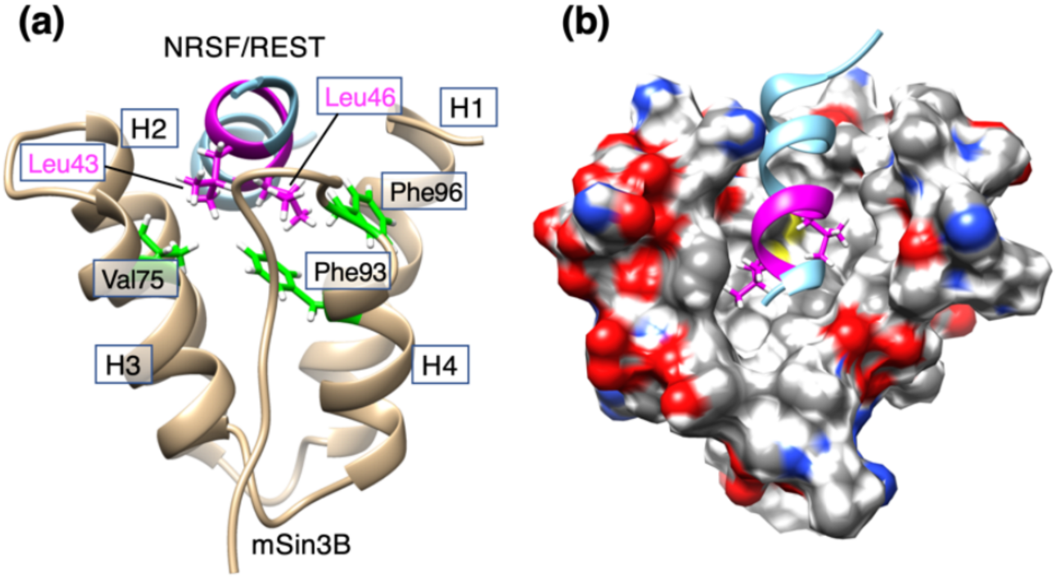

Figure 4-11. NRSF/REST–mSin3B complex structure determined by NMR (PDB ID: 2CZY). (a) The LIML sequence of NRSF/REST is shown by magenta-colored ribbon model, and two hydrophobic residues Leu 43 and Leu 46 (magenta-colored labels) of the LIML sequence are displayed explicitly. These two residues contact to three hydrophobic residues Val 75, Phe 93, and Phe 96 (green-colored residues with black labels) of mSin3B. Labels H1, ..., H4 indicate helices 1, ..., 4 of mSin3B (PAH1 domain), respectively. (b) mSin3B is presented by a surface model, where hydrophobic surface is represented by white color.

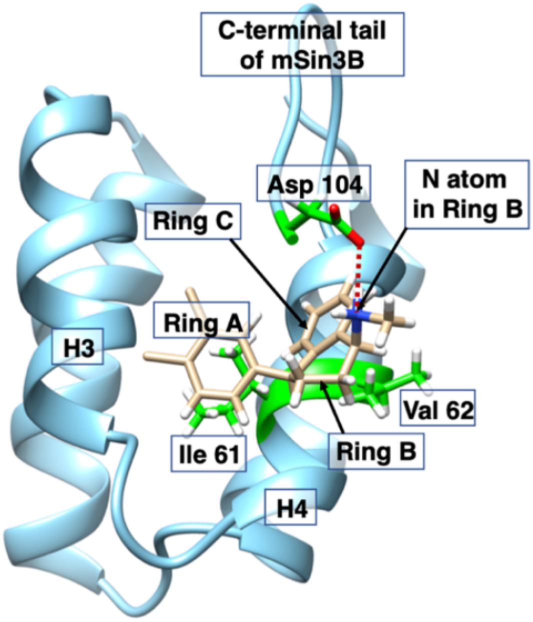

Figure 4-12. Conformation of sertraline taken from Cluster B. See Figure 4-10a for definition of Cluster B. The nitrogen atom of sertraline's Ring B interacts electrostatically to an oxygen atom of Asp 104 of the C-terminal tail of mSin3B, which is shown by brown-colored broken line. The distance between the two atom is 2.8 Å. Hydrophobic contacts are formed between sertraline's Ring A and Ile 61 of helix H4 of mSin3B as well as between sertraline's Ring C and Val 62.

4.3.3 Radial distribution function of ligand-cleft distance

Figure 4-10 is insightful to guess the ligand–mSin3B interaction. Nevertheless, $\rho_{CG}^{(s)}(\boldsymbol{r})$ was calculated using the geometrical centers of the ligands. There is a possibility that some parts of the ligands contacted to the hydrophobic cleft tightly even when the geometrical center was distant from the cleft. For instance, acitretin is a long molecule, and then the head or tail of acitretin (Figure 4-2c) may be inserted to the cleft keeping the geometrical center out of the cleft.

To make clear this point, I calculated a radial distribution function (RDF) between the ligand and three residues Val 75, Phe 93, and Phe 96 located in the mSin3B cleft.

Defining a distance r between two objects in a system, we can calculate a distance distribution function (DDF) $P(r)$, where the probability of detecting r in a window $[r, r + dr]$ is $P(r)dr$. The normalization of DDF is defined as $\int_0^\infty P(r)dr = 1$. I use w_1 (section 4.2.5) to calculate the DDF. Then, the radial distribution function (RDF) $p(r)$ is defined formally as: $p(r) = P(r)/4\pi r^2$. This function is normalized as $\int_0^\infty p(r)4\pi r^2 dr = 1$.

In the present study, I calculate the minimum heavy-atomic distance $r_R^{(s)}$ between the ligand and one of three residues Val 75, Phe 93, and Phe 96 in the mSin3B cleft, where the subscript R of $r_R^{(s)}$ is the residue specifier (R = Val 75, Phe 93, or Phe

96) and the superscript s is the system specifier defined in eq. (4.2). Then, I calculated three RDFs $p\left(r_R^{(s)}\right)$ for each system.

Figure 4-13 demonstrates the resultant RDFs. For all of the systems, the function $p(r_{Phe96}^{(s)})$ exhibited a peak at 3.5 Å regardless of the peak height. This result is plausible because the residue Phe 96 is located at the entrance of the cleft (Figure 4-11a): The ligand can contact to Phe 96 without sinking into the cleft. More important RDFs are $p(r_{Val75}^{(s)})$ and $p(r_{Phe93}^{(s)})$ because Val 75 and Phe 93 are located at the bottom of the cleft. Apparently, the peaks of the RDFs at $r_R^{(s)}{\sim}4$ Å for the acitretin–mSin3B system were considerably smaller than those for sertraline–mSin3B and YN3–mSin3B systems. This means that acitretin did not interact frequently or tightly with the bottom of the cleft. The highest peaks were from the sertraline-mSin3B system (Fig. 4-13a), and the peaks from the YN3–mSin3B were intermediate between sertraline and acitretin (Fig. 4-13b). These results suggest that sertraline may resemble the ligand–receptor interaction mode found in the NRSF/REST–mSin3B complex.

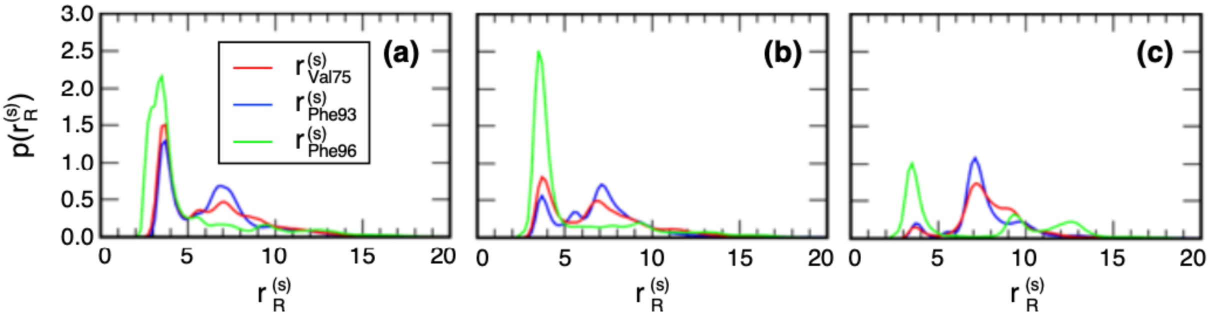

Figure 4-13. Radial distribution functions (RDFs) $p(r_R^{(s)})$ for (a) the sertraline–mSin3B (s = sertraline), (b) YN3–mSin3B (s = YN3), and (c) acitretin–mSin3B systems (s = acitretin). Three RDFs $p(r_{Val75}^{(s)})$, $p(r_{Phe93}^{(s)})$, and $p(r_{Phe96}^{(s)})$ are shown by different colors as indicated in insets of panels.

To investigate more the ligand–receptor interactions for the sertraline–mSin3B system, I calculated the spatial density $\rho_{GCA}^{(\text{sertraline})}(\boldsymbol{r})$ of the geometrical center of Ring A of sertraline: See Figure 4-14a for the position of Ring A and eq. (4.4) for the procedure to calculate $\rho_{GCA}^{(s)}(\boldsymbol{r})$. Figure 4-4b illustrates $\rho_{GCA}^{(s)}(\boldsymbol{r})$ and $\rho_{GC}^{(s)}(\boldsymbol{r})$ with the contour level of $0.5\rho_0$. The contours in the mSin3B's cleft correspond those for Cluster A in Fig. 4-10a. Thus, conformations taken from the contours are constituents of the most probable binding mode of the sertraline–mSin3B system. Ring A occupies the deeper position of the cleft than the geometrical center of the entire sertraline did. Thus, Ring A was closer to Val 75, Phe 93, and Phe 96 of mSin3B. In the next subsection, I investigate the conformations in this most probable binding mode. The contours near the C-terminal tail of mSin3B does to Cluster B, which is the artificial cluster as discussed before.

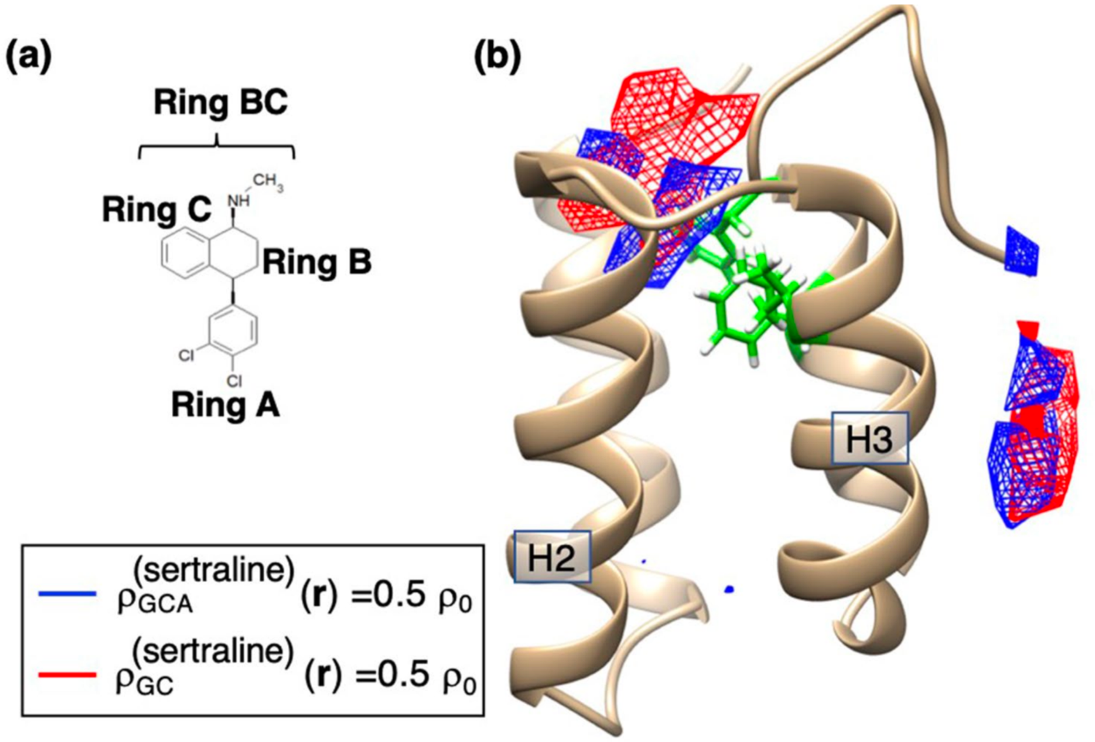
(a)
Ring BC
Ring C
NH
CH3
Ring B
Cl
Cl
Ring A
(b)
H3
H2
$\rho_{GCA}^{(\text{sertraline})}(\mathbf{r}) = 0.5\ \rho_0$
$\rho_{GC}^{(\text{sertraline})}(\mathbf{r}) = 0.5\ \rho_0$

Figure 4-14. (a) Chemical structure of sertraline, where Ring A and Ring BC (Rings B and C) are shown. (b) Spatial density $\rho_{CG}^{(sertraline)}(\boldsymbol{r})$ (blue-colored contours) for the geometric center of Ring A of sertraline in the sertraline–mSin3B system, where the contour level is $\rho_{CGA}^{(sertraline)}(\boldsymbol{r}) > 0.5\rho_0$ ($\rho_0 = 0.001$). Red-colored contours are $\rho_{CG}^{(sertraline)}(\boldsymbol{r})$, which is spatial density of the geometric center of the entire sertraline. Labels H2 and H3 represent helices 2 and 3, respectively. Green-colored residues are Val 75, Phe 93, and Phe 96 of mSin3B. See Section 4.2.5 for procedure to calculate $\rho_{CGA}^{(sertraline)}(\boldsymbol{r})$.

4.3.4 Orientation of ligands bound to mSin3B

The above analysis shows that the most stable/probable complex is not assigned to a single complex structure but to an ensemble of complex structures, which construct a high-density cluster. For the sertraline–mSin3B system, the most probable complex structures are those originated from Cluster A (Fig. 4-10a). Figure 4-15 demonstrates some conformations picked randomly from Cluster A. This figure also displays Val 75, Phe 93, and Phe 96 of mSin3B. Here I judged that Ring A was contacting to these residues when $r_R^{(s)} < 4.0$ Å was satisfied. Remember that $r_R^{(s)}$ was defined in the previous subsection. In Fig. 4-15a Ring A contacted to Val 75 and Phe 93, and in Fig. 4-15b Ring A did to Phe 93 and Phe 96. Ring A contacted to Phe 96 in Fig. 4-15c, and Ring A did to all of the three residues in Fig. 4-15d. Examining a number of snapshots, I

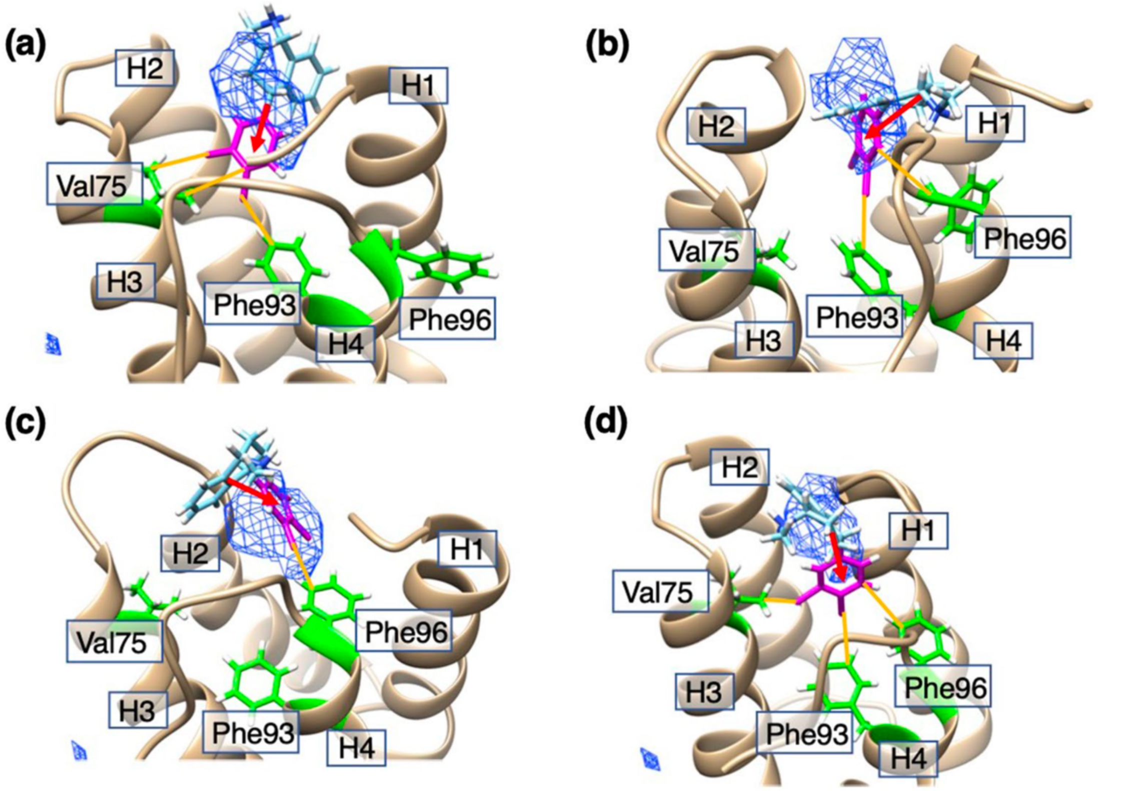

Figure 4-15. (a)–(d) Sertraline–mSin3B complexes picked randomly from the high-density cluster (Cluster A) of $\rho_{CG}^{(sertraline)}(\boldsymbol{r}) > 0.5\rho_0$ (ρ_0 = 0.001), which are shown by blue-colored contours. Magenta-colored portion is Ring A of sertraline. Red-colored arrows are those pointing from the Ring-BC geometrical center to that of Ring A (see Fig. 4-16a for positions of Ring A and BC). Green-colored residues are Val 75, Phe 93, and Phe 96 of mSin3B. Ocher-colored lines represent contacts between sertraline and the three residues. Labels H1, ..., H4 are helices 1–4 of mSin3B (PAH1 domain).

concluded that ligand–receptor contacts shown in Fig. 4-13a were contributed mainly by Ring A.

Red-colored arrows in Fig.4-15 shows the molecular orientations of sertraline has a tendency: Ring A was inserted into the binding cleft of mSin3B and Ring BC was left behind. To analyze more this tendency, I defined the molecular orientation of a snapshot for the three ligands: Fig. 4-16a, d, and g illustrate the ligand's orientation for sertraline, YN3, and acitretin, respectively.

Molecular orientation vectors for the compounds are defined by the red-colored vectors in Figures 4-16a, 4-16d, and 4-16g. Then, I defined a unit vector parallel to the molecular orientation vector for each snapshot: $\boldsymbol{e}_i^{(s)}$ where the subscript i is the snapshot specifier and superscript s is the system specifier (s = sertraline, YN3, or acitretin). Then, the average of $\boldsymbol{e}_i^{(s)}$ at each cube was defined as:

$$\langle \boldsymbol{e}^{(s)}(\boldsymbol{r}) \rangle = \frac{1}{\sum_i w_i \delta_{GC}(\boldsymbol{r}; i)} \sum\nolimits_i \boldsymbol{e}_i^{(s)} w_i \delta_{GC}(\boldsymbol{r}; i). \tag{4.6}$$

I calculated $\langle \boldsymbol{e}^{(s)}(\boldsymbol{r}) \rangle$ in cubes whose densities were $\rho_{GC}^{(s)}(\boldsymbol{r}) \geq 0.5\rho_0$ for sertraline and YN3, and $\rho_{GC}^{(s)}(\boldsymbol{r}) \geq 0.1\rho_0$ for acitretin. I aim to investigate the spatial pattern of $\langle \boldsymbol{e}^{(s)}(\boldsymbol{r}) \rangle$ in the high-density regions. Thus, I set the threshold for sertraline and YN3 was set to $0.5\rho_0$ referring to Figs. 4-10a and 4-10b. On the other hand, the regions of $0.5\rho_0$ did not appear around the cleft of mSin3B substantially for acitretin (Fig. 4-10c). Thus, I decreased the threshold to $0.1\rho_0$ for acitretin.

Note that $|\langle \boldsymbol{e}^{(s)}(\boldsymbol{r}) \rangle|$ represents a degree of ligand orientational ordering in each cube. $|\langle \boldsymbol{e}^{(s)}(\boldsymbol{r}) \rangle|$ takes the maximum of 1 when all $\boldsymbol{e}_i^{(s)}$ have exactly the same orientations in the cube. If snapshots have uncorrelated orientations, then $|\langle \boldsymbol{e}^{(s)}(\boldsymbol{r}) \rangle|$ becomes small.

Figure 4-16b and c demonstrate the spatial patterns of $\langle \boldsymbol{e}^{(sertraline)}(\boldsymbol{r}) \rangle$ for the sertraline–mSin3B system viewed from two different directions. This figure also indicates that Ring A tends to be inserted in the cleft of mSin3B, which is consistent to Fig. 4-15.

I also calculated $\langle \boldsymbol{e}^{(YN3)}(\boldsymbol{r}) \rangle$ for the YN3–mSin3B (Fig. 4-16e and f) and acitretin–mSin3B (Fig. 4-16h and i) systems. Figure 4-16h and i show that the acitretin's orientation tends to be parallel or anti-parallel to the helical cylinder of NRSF/REST in the NRSF/REST–mSin3B complex. I presume that these acitretin's orientations have an advantage to fit the whole acitretin's framework to the binding cleft of mSin3B. To stabilize one of the parallel or anti-parallel orientations of acitretin, an additional inter-molecular interaction is required, which works differently between the two orientations.

I consider that there is no such interaction to stabilize effectively one of the two orientations for the acitretin–mSin3B system.

On the other hand, the YN3's orientations (Fig. 4-16e and f) were not aligned to the helical cylinder of REST/ NRSF but tilting from the helical cylinder. I presume that this tilting of YN3 is because YN3 is smaller than acitretin: YN3 may be responding to undulations of mSin3B's molecular surface in the cleft. This point is discussed later again.

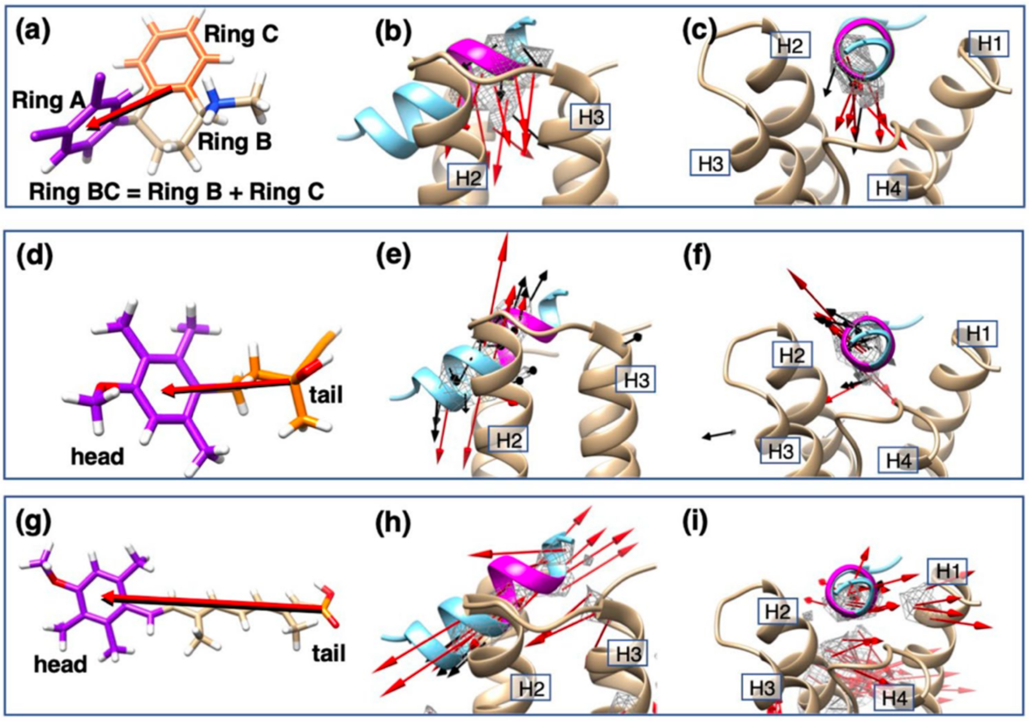

Figure 4-16. (a) A snapshot of sertraline, where three rings, Ring A, B, and C are defined and Rings B and C are unified as Ring BC. Ligand's orientation (red-colored arrow) is defined by an arrow pointing from the geometric center of Ring BC to that of Ring A. Panels (b) and (c) illustrate spatial patterns of $\langle \boldsymbol{e}^{(sertraline)}(\boldsymbol{r}) \rangle$ in the cleft of mSin3B viewed from two different directions. The vectors $\langle \boldsymbol{e}^{(sertraline)}(\boldsymbol{r}) \rangle$ are assigned to high-density regions of $\rho_{CG}^{(sertraline)}(\boldsymbol{r}) > 0.5\rho_0$ ($\rho_0 = 0.001$) presented by gray contours. Red and black vectors are those with $\left|\langle \boldsymbol{e}^{(sertraline)}(\boldsymbol{r}) \rangle\right| \geq 0.5$ and $\left|\langle \boldsymbol{e}^{(sertraline)}(\boldsymbol{r}) \rangle\right| < 0.5$, respectively. This figure also displays NRSF/REST bound to mSin3B (PDB ID: 2CZY), and the magenta-colored segment is the LIML sequence of NRSF/REST. Note that NRSF/REST is not involved in the current simulation. Labels H1, ..., H4 are helices 1, ..., 4 of mSin3B (PAH1 domain). (d) A snapshot of YN3, where "head" and "tail" are colored by purple and orange, respectively. Red-colored arrow is the YN3's orientation pointing from the geometric center of tail to that of head. Panels (e) and (f) illustrate spatial patterns of $\langle \boldsymbol{e}^{(YN3)}(\boldsymbol{r}) \rangle$. See captions for panels (b) and (c) for method to draw $\langle \boldsymbol{e}^{(YN3)}(\boldsymbol{r}) \rangle$ from two different directions. (g) A snapshot of acitretin, where "head" and "tail" are colored by purple and orange, respectively. Red-colored arrow is the acitretin's orientation pointing from the geometric center of tail to that of head. Panels (h) and (i) illustrate spatial patterns of $\langle \boldsymbol{e}^{(acitretine)}(\boldsymbol{r}) \rangle$ from two different directions. See captions for panels (b) and (c) for method to draw $\langle \boldsymbol{e}^{(acitretine)}(\boldsymbol{r}) \rangle$.

4.3.5 Flexibility of ligands' framework

Figure 4-15 demonstrates the structural variety of the sertraline's framework in the most probable complex state (Cluster A). To quantify the flexibility of the ligand's framework, I calculated the distance distribution function, DDF, for a distance between two atoms set in each ligand. The procedure to calculate DDF for a distance is given in Section 4.3.2.

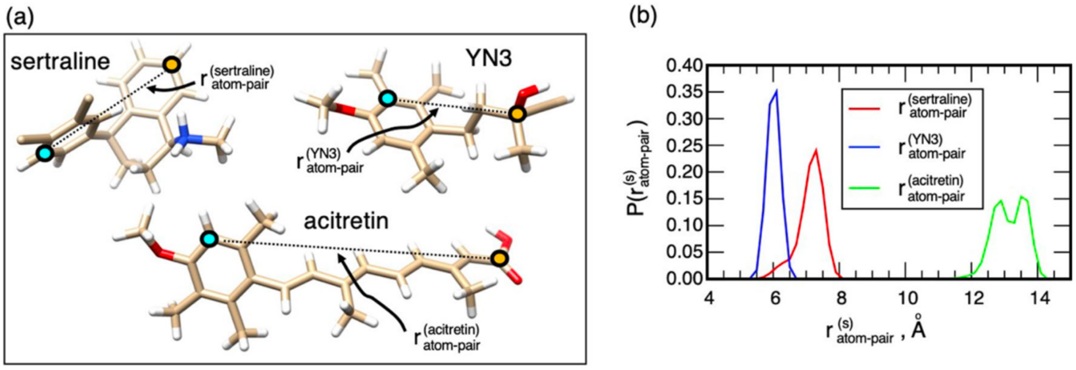

Figure 4-17. (a) Structures of the three compounds sertraline, YN3, and acitretin. For each compound, inter-atomic distance $r^{(s)}_{atom-pair}$ (s = sertraline, YN3, or acitretin) is defined between cyan- to orange-colored atoms. A cyan-colored atom, which is involved in a ring for each ligand, is not set on the ring rotation axis to detect the ring-rotational motions. (b) Distance distribution functions (DDFs), $P(r^{(s)}_{atom-pair})$, for the three distances.

Here, I picked the cyan- and orange-colored atoms in Figure 4-17a from the framework of each ligand, and calculated the distance between the two atoms: $r^{(s)}_{atom-pair}$. The DDFs $P(r^{(s)}_{atom-pair})$ for these distances are displayed in Figure 4-17b. The average of the distance $< r^{(s)}_{atom-pair} >$ and its standard deviation (amount of fluctuations) $SD(r^{(s)}_{atom-pair})$ were: 7.12 Å and 0.37 Å for sertraline, 6.01 Å and 0.14 Å for YN3, and 13.15 and 0.47 Å for acitretin. The largest SD was assigned to acitretin. However, this does not mean that acitretin is the most flexible, because a long molecule may have generally a large SD even if the ligand is stiff, and because acitretin is the longest ligand of the three. I consider that the amount of fluctuations per unit length, $sd^{(s)} = SD(r^{(s)}_{atom-pair})/< r^{(s)}_{atom-pair} >$, is a better quantity to quantify the flexibility of the molecular flexibility. The resultant $sd^{(s)}$ was 5.20×10^{-2} for sertraline, 2.33×10^{-2} for YN3, and 3.57×10^{-2} for acitretin. From these values, the framework of sertraline is the most flexible, acitretin has a considerably stiffer framework than sertraline does, and YN3 is stiffer than acitretin. The flexibility of the framework may be related the ligand's molecular orientation as discussed later.

4.4 Discussion

Starting from the initial conformations where the ligands were distant from the cleft of mSin3B (Figure 4-5), only sertraline reproduced a similar intermolecular hydrophobic core with that observed in the REST/ NRSF–mSin3B complex (PDB ID: 2CZY): Ring A of sertraline, which is hydrophobic, bound deeply to the cleft of mSin3B with contacting to the hydrophobic sidechains sited in the cleft (Fig. 4-15). The binding scheme for the sertraline–mSin3B complex is summarized in the column of Fig. 4-18. The hydrophobic core formed between Ring A of sertraline and the cleft of mSin3B was similar with that formed in the NRSF/REST–mSin3B complex. I presume that the sertraline–mSin3B complex formation competes with the NRSF/REST–mSin3B complex formation, and that this competition leads to the medulloblastoma cell-growth inhibition activity of sertraline.

It is interesting to compare the currently obtained complex structures (i.e., structures in Cluster A) with a modeled structure[82], which was obtained from a docking software HADDOCK[96]. Although HADDOCK does not take the statistical-mechanical factors into the modeling, the receptor and ligand are modeled to bind to each other so as to satisfy experimental data, the chemical shift perturbation data[82]. Figure 4-19 displays the HADDOCK complex structure using atomic coordinates presented by Kurita et al. Interestingly, Ring A sertraline in the HADDOCK model contacted to Phe 93 and Phe 96 (the ocher-colored lines in the figure), and then, Ring A was oriented somewhat toward the inside of the mSin3B's cleft (the red-colored arrow in the figure). Remember that these structural features were found in our computed complex structures in Cluster A (Figs. 4-15 and 4-16a). I emphasize that the current simulation did not use the experimental data in computation.

Now, I discuss the interactions of acitretin with mSin3B. Acitretin has a low spatial density in the mSin3B's cleft (Fig. 4-10c), which weakens the inhibitory activity partly. Furthermore, the framework of acitretin tends to be parallel or anti-parallel to the mSin3B's cleft (Fig. 4-16c and the right panel of Fig. 4-18b). Remember that acitretin can be regarded as a long and stiff rod as shown in the above section. The parallel or anti-parallel molecular orientation is advantageous to be fit to the cleft. If acitretin has a flexible framework, acitretin may insert a portion into the hydrophobic cleft by bending the framework.

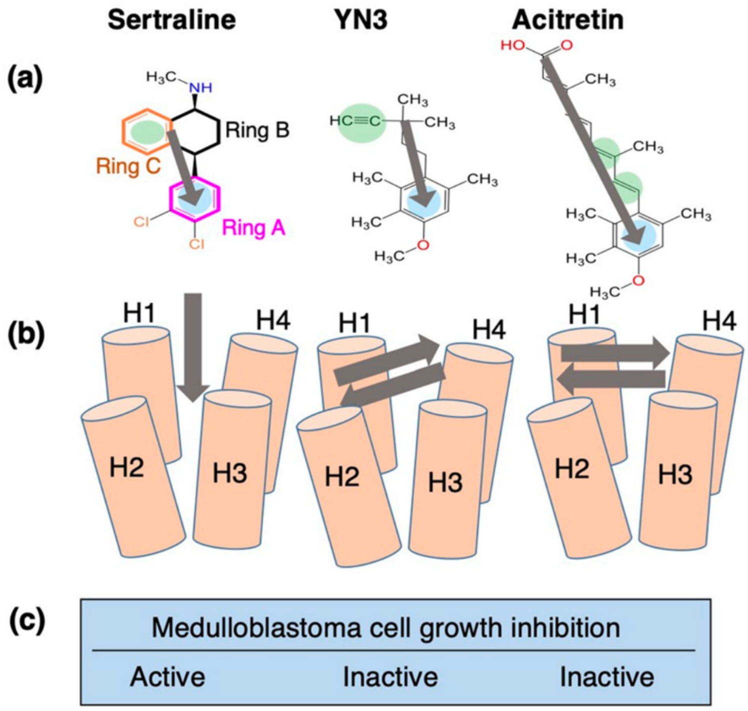

Figure 4-18. (a) Chemical structures of three compounds, and their orientations (gray arrows). Rings A, B and C of sertraline are depicted in magenta, black, and orange, respectively. Blue circles represent aromatic rings that are expected to interact with the hydrophobic cleft of mSin3B from conventional structure–activity relationship (SAR). Green circles represent π-electron rich regions. (b) Orientations of the compounds in the bound forms with mSin3B, resulted from the present MD simulation study. (c) Presence/absence of medulloblastoma cell-growth inhibition activities for the compounds.

Next, I discuss the interactions of YN3 with mSin3B. Although YN3 had a high spatial density around the mSin3B's cleft, YN3 did not sink deeply in the cleft and the hydrophobic core was not formed (Fig. 4-10b). This result is natural because YN3 can be regarded as a stiff rod. As discussed for acitretin above, the stiff rod fits well to the cleft taking the parallel or anti-parallel orientation. On the other hand, I presume that the slight tilt of $<\boldsymbol{e}^{(YN3)}(\boldsymbol{r})>$ to the helix cylinder of NRSF/REST in the NRSF/REST–mSin3B complex (Fig. 4-16b and the middle panel of Fig. 4-18b) is resulted from the small size of YN3. I.e., YN3 can adopt to the jaggedness of the inside of the cleft. If YN3 has more flexibility, YN3 may exert the inhibitory activity by inserting a molecular portion into the cleft with varying the molecular conformation. On the other hand, the added flexibility may induce binding of YN3 to the other surfaces of mSin3B than the cleft. These competitive effects of the molecular flexibility cannot be assessed only from the chemical structure of the compound.

I also note another chemical property of YN3 and acitretin, which is disadvantageous for interacting with hydrophobic cleft of mSin3B: Both rings in these compounds have an oxygen atom (Figure 4-2b and c), which can interact to hydrophilic residues in mSin3B or water molecules electrostatically or by hydrogen bonding.

The conventional structure–activity relationship (SAR) study has been based on the chemical structures and target-binding activity data. Figure 4-18a shows one of the possible alignments of the three compounds based on their chemical structures. Each compound has an aromatic ring corresponding to Ring A and each aromatic ring has side chains in the para position. While sertraline has Cl atoms on Ring A, YN3 and acitretin have methoxy groups whose volumes are close to that of the Cl atom. Instead of Ring C of sertraline, YN3 and acitretin have triple and double bonds, respectively. These bonds are π-electron rich, and can form CH–π interaction as same as aromatic rings. However, such SAR analysis could not explain presence and absence of the medulloblastoma cell-growth inhibition activities of their compounds (Fig. 4-18c). The present MD simulation study predicted the molecular orientations of the compounds (Fig. 4-18b) and the presence/absence of the hydrophobic core in the cleft of mSin3B (Fig. 4-10). Importantly, only sertraline could reproduce the binding mode in the NRSF/REST–mSin3B complex. I emphasize that the current simulation method, the GA-guided mD-VcMD simulation, produces a thermodynamically acceptable ensemble consisting of various conformations (bound and unbound conformations), and importantly a thermodynamic weight is assigned to each snapshot in the ensemble.

The preceding study[82] classified 52 compounds into two pharmacophores, A and B, based on their chemical structures (see Fig. 2 of Ref.82), where sertraline belongs to Pharmacophore A and the YN3 to Pharmacophore B. Because acitretin has a structural similarity with YN3 apparently, acitretin belongs to Pharmacophore B. Based on Figs. 4-16 and 4-18b, the compounds belonging to Pharmacophore B have the parallel or anti-parallel orientation, and the compound belonging to Pharmacophore A has a perpendicular orientation. Therefore, the current MD procedure is useful if it is used with the pharmacophore analysis.

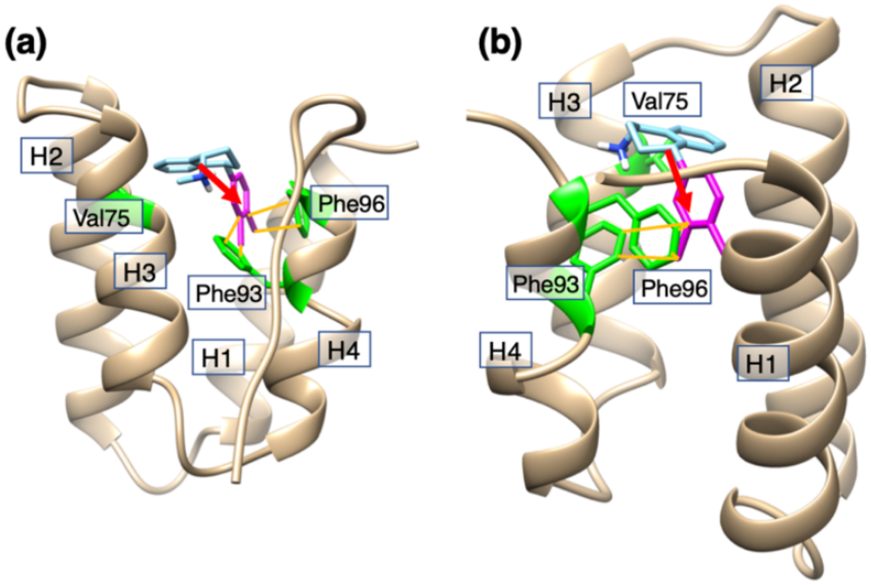

Figure 4-19. Sertraline–mSin3B complex viewed from two different orientations (a) and (b), which was modeled by Kurita et al. [82] using HADDOCK modeling[96]. The atomic coordinates were provided from Kurita et al. In the HADDOCK modeling, the receptor and ligand bind to each other so as to satisfy chemical shift perturbation data. The magenta-colored portion is Ring A of sertraline. Red-colored arrow is the ligand's orientation vector pointing from the Ring-BC geometrical center to that of Ring A. Green-colored residues are Val 75, Phe 93, and Phe 96 of mSin3B. Ocher-colored lines represent contacts between sertraline and the three residues. Labels H1, ..., H4 are helices 1-4 of mSin3B (PAH1 domain)

4.5 Conclusions

Binding of the compounds to mSin3B (PAH1 domain) was investigated by the GA-guided mD-VcMD simulation. This method produced useful quantities such as the spatial density of the ligand around the receptor (Fig. 4-10), the intermolecular contact patterns (Fig. 4-13), the propensity of molecular orientation (Figs. 4-15 and 4-16), and the ligand flexibility (Fig. 4-17). From these analyses, I showed that only sertraline produces a similar inter-molecular binding mode observed in the NRSF/REST–mSin3B complex. Figure 4-20 is a schematic drawing to design an inhibitor. Given a framework of the compound, by adding a hydrophobic sidechain to the framework, the hydrophobic core is formed between the sidechain and the hydrophobic cleft of mSin3B. The flexibility of the compound's framework may increase the binding affinity, although a long and stiff framework may decrease the binding affinity. In general, it is difficult to specify the effect of the modification to the biological activity (i.e., inhibitory activity in the present study) only from the compound's chemical structure. In contrast, the GA-guided mD-VcMD is useful to identify the effect with analyzing the thermally equilibrated conformational ensemble.

Because NRSF/REST is an intrinsically disordered segment that can bind to multiple proteins[55,56] as mentioned in "4.1 Introduction" section, the current study is an example to design a compound that can inhibit binding of an intrinsically disorder segment to a protein receptor.

The 3D models have been submitted to the Biological Structure Model Archive (BSM-Arc) of the PDBj under BSM-ID BSM00020 (https://bsma.pdbj.org/entry/20), which are freely available[97].

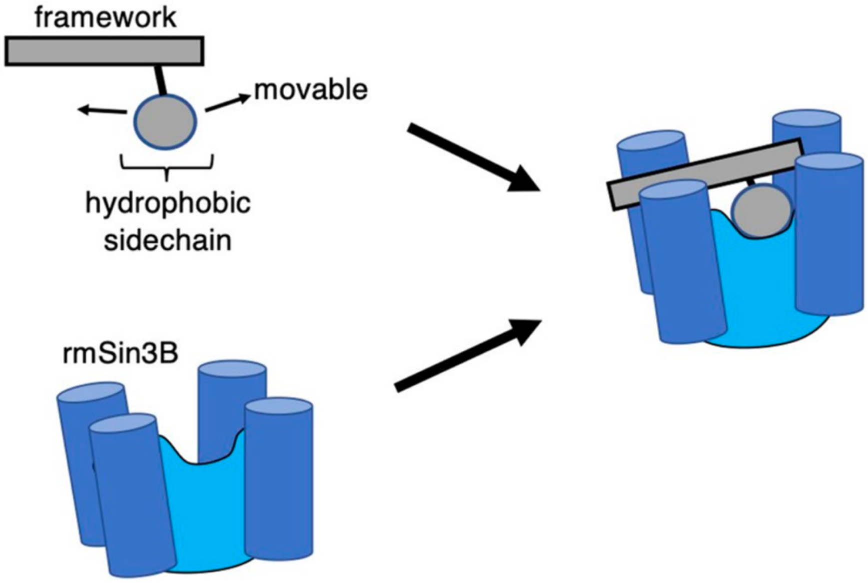

Figure 4-20. Scheme of a compound that binds to the cleft of mSin3B.

Chapter V

General Conclusions

In this study, I introduced the multidimensional virtual-system coupled canonical molecular dynamics (mD-VcMD) method to compute free-energy landscapes of protein–protein and protein–ligand interaction (Chapter II). First, I evaluated the method using simple molecular models consisting of three and four alanine-peptides in explicitly solvated systems (Chapter III). Second, I investigated the spatial distribution of three compounds sertraline, YN3, and acitretin, respectively, around the PAH1 domain of mSin3B, which are obtained from the GA-guided mD-VcMD simulation of the systems (Chapter IV).

In Chapter III, before applying the mD-VcMD method to complicated systems to resolve difficulties related to biology, I investigated this method's capacity to build correct canonical ensembles using the simple systems including consisting of alanine peptides. The resultant ensembles of 2D-VcMD and 3D-VcMD for 3-ALA and 4-ALA systems were well converged to the results of long-term conventional MD simulation (Figs 3-5, 3-6, and 3-7). And the conformational ensembles covered various contacting topologies of the three and four peptides (Figs 3-10, 3-11, and 3-12).

These results show the mD-VcMD method is useful to investigate molecule interactions using computer simulation according to the arbitrary reaction coordinates. Although I presented applications only for simple molecular systems with weak interactions, mD-VcMD is applicable for systems involving stable molecular interaction in which a ligand binds to a deep pocket of a receptor.

In Chapter IV, the binding of the compounds to mSin3B (PAH1 domain) was investigated by the GA-guided mD-VcMD simulation. This method produced useful quantities such as the spatial density of the ligand around the receptor (Fig. 4-10), the intermolecular contact patterns (Fig. 4-13), the propensity of molecular orientation (Figs. 4-15 and 4-16), and the ligand flexibility (Fig. 4-17).

The results of simulations for the three systems consisting of mSin3B and one of three compounds are briefly described below.

For sertraline and YN3, a high-density conformational cluster (designated Cluster A in Fig. 4-10) can be found in the cleft of mSin3B, but acitretin did not display a notable cluster in the cleft. This suggests that ligand–mSin3B binding is stronger for sertraline and YN3 than for acitretin.

According to Fig. 4-13, the RDF peaks at $r_R^{(s)} \sim 4$ Å for the acitretin–mSin3B system were much lower than those for the sertraline–mSin3B and YN3–mSin3B systems, indicating that the acitretin did not interact with the bottom of the cleft frequently or tightly. The highest peaks of the RDFs were from the sertraline–mSin3B system (Fig. 4-13a), and the peaks from the YN3–mSin3B were intermediate between sertraline and acitretin (Fig. 4-13b).

Figure 4-15 and 4-16 show the following three points related to the molecular orientations of compounds; (1) The molecular orientations of sertraline has a tendency: Ring A of sertraline was placed into the binding cleft of mSin3B, whereas Ring BC remained outside. (2) The acitretin's orientation tends to be parallel or anti-parallel to the helical cylinder of NRSF/REST in the NRSF/REST–mSin3B complex. (3) The orientations of the YN3 were tilting from the helical cylinder of NRSF/REST rather than aligned to it.

Figure 4-17 shows that the framework of sertraline is the most flexible, acitretin has a considerably stiffer framework than sertraline does, and YN3 is stiffer than acitretin.

From these analyses, I discussed that only sertraline produces a similar inter-molecular binding mode observed in the NRSF/REST–mSin3B complex. In general, it is difficult to specify the effect of the modification to the biological activity (i.e., inhibitory activity in the present study) only from the compound's chemical structure. In contrast, the GA-guided mD-VcMD is useful to identify the effect with analyzing the thermally equilibrated conformational ensemble.

Finding a proper definition of a collection of reaction coordinates is a major issue for the mD-VcMD approach as well as other existing methods, such umbrella sampling, when addressing complicated systems. However, by discretizing the reaction-coordinate space and biasing with the flat-bottom potential, higher-dimensional reaction-coordinate spaces can be applied more easily. Even if finding the optimal definition of a single reaction coordinate is difficult, temporarily applying some multiple reaction coordinates and finding the optimal coordinates, which can be constructed by linear combinations of the previously introduced ones, may provide a practical solution.

Bibliography

1. Berg, B. A. & Neuhaus, T. Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Phys Rev Lett* **68**, 9–12 (1992).
2. Lee, J. New Monte Carlo algorithm: Entropic sampling. *Phys Rev Lett* **71**, 211 (1993).
3. Hukushima, K. & Nemoto, K. Exchange Monte Carlo Method and Application to Spin Glass Simulations. *http://dx.doi.org/10.1143/JPSJ.65.1604* **65**, 1604–1608 (2013).
4. Swendsen, R. H. & Wang, J. S. Replica Monte Carlo Simulation of Spin-Glasses. *Phys Rev Lett* **57**, 2607 (1986).
5. Hansmann, U. H. E. & Okamoto, Y. Prediction of peptide conformation by multicanonical algorithm: New approach to the multiple-minima problem. *J Comput Chem* **14**, 1333–1338 (1993).
6. Hansmann, U. H. E., Okamoto, Y. & Eisenmenger, F. Molecular dynamics, Langevin and hydrid Monte Carlo simulations in a multicanonical ensemble. *Chem Phys Lett* **259**, 321–330 (1996).
7. Kidera, A. Enhanced conformational sampling in Monte Carlo simulations of proteins: application to a constrained peptide. *Proc Natl Acad Sci U S A* **92**, 9886–9889 (1995).
8. Iba, Y., Chikenji, G. & Kikuchi, M. Simulation of Lattice Polymers with Multi-Self-Overlap Ensemble. *J Physical Soc Japan* **67**, 3327–3330 (1998).
9. Nakajima, N., Nakamura, H. & Kidera, A. Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. *Journal of Physical Chemistry B* **101**, 817–824 (1997).
10. Wang, F. & Landau, D. P. Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. *Phys Rev Lett* **86**, 2050 (2001).
11. Wang, F. & Landau, D. P. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Phys Rev E* **64**, 056101 (2001).
12. Darve, E. & Pohorille, A. Calculating free energies using average force. *J Chem Phys* **115**, 9169 (2001).
13. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **99**, 12562–12566 (2002).

14. Martoňák, R., Laio, A. & Parrinello, M. Predicting Crystal Structures: The Parrinello-Rahman Method Revisited. *Phys Rev Lett* **90**, 4 (2003).

15. Kim, J. G., Fukunishi, Y. & Nakamura, H. Multicanonical molecular dynamics algorithm employing an adaptive force-biased iteration scheme. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* **70**, 4 (2004).

16. Nagasima, T., Kinjo, A. R., Mitsui, T. & Nishikawa, K. Wang-Landau molecular dynamics technique to search for low-energy conformational space of proteins. *Phys Rev E Stat Nonlin Soft Matter Phys* **75**, 066706 (2007).

17. Hori, N., Chikenji, G., Berry, R. S. & Takada, S. Folding energy landscape and network dynamics of small globular proteins. *Proc Natl Acad Sci U S A* **106**, 73–78 (2009).

18. Higo, J., Nishimura, Y. & Nakamura, H. A free-energy landscape for coupled folding and binding of an intrinsically disordered protein in explicit solvent from detailed all-atom computations. *J Am Chem Soc* **133**, 10448–10458 (2011).

19. Ikebe, J., Sakuraba, S. & Kono, H. Adaptive lambda square dynamics simulation: An efficient conformational sampling method for biomolecules. *J Comput Chem* **35**, 39–50 (2014).

20. Kasahara, K., Shiina, M., Higo, J., Ogata, K. & Nakamura, H. Phosphorylation of an intrinsically disordered region of Ets1 shifts a multi-modal interaction ensemble to an auto-inhibitory state. *Nucleic Acids Res* **46**, 2243–2251 (2018).

21. Mitsutake, A., Sugita, Y. & Okamoto, Y. Generalized-Ensemble Algorithms for Molecular Simulations of Biopolymers. (2001) doi:10.1002/1097-0282.

22. Higo, J., Ikebe, J., Kamiya, N. & Nakamura, H. Enhanced and effective conformational sampling of protein molecular systems for their free energy landscapes. *Biophys Rev* **4**, 27–44 (2012).

23. Higo, J. *et al.* Virtual-system-coupled adaptive umbrella sampling to compute free-energy landscape for flexible molecular docking. *J Comput Chem* **36**, 1489–1501 (2015).

24. Higo, J., Kasahara, K., Dasgupta, B. & Nakamura, H. Enhancement of canonical sampling by virtual-state transitions. *J Chem Phys* **146**, 044104 (2017).

25. Higo, J., Kasahara, K. & Nakamura, H. Multi-dimensional virtual system introduced to enhance canonical sampling. *J Chem Phys* **147**, 134102 (2017).

26. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J Comput Phys* **23**, 187–199 (1977).

27. Torrie, G. M. & Valleau, J. P. Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid. *Chem*

Phys Lett **28**, 578–581 (1974).

28. Mezei, M. Adaptive umbrella sampling: Self-consistent determination of the non-Boltzmann bias. *J Comput Phys* **68**, 237–248 (1987).

29. Babin, V., Karpusenka, V., Moradi, M., Roland, C. & Sagui, C. Adaptively biased molecular dynamics: An umbrella sampling method with a time-dependent potential. *Int J Quantum Chem* **109**, 3666–3678 (2009).

30. Wojtas-Niziurski, W., Meng, Y., Roux, B. & Bernèche, S. Self-learning adaptive umbrella sampling method for the determination of free energy landscapes in multiple dimensions. *J Chem Theory Comput* **9**, 1885–1895 (2013).

31. Dasgupta, B., Nakamura, H. & Higo, J. Flexible binding simulation by a novel and improved version of virtual-system coupled adaptive umbrella sampling. *Chem Phys Lett* **662**, 327–332 (2016).

32. Huber, T., Torda, A. E. & van Gunsteren, W. F. Local elevation: A method for improving the searching properties of molecular dynamics simulation. *J Comput Aided Mol Des* **8**, 695–708 (1994).

33. Bieler, N. S. & Hünenberger, P. H. Communication: Estimating the initial biasing potential for λ-local-elevation umbrella-sampling (λ-LEUS) simulations via slow growth. *J Chem Phys* **141**, 201101 (2014).

34. Hansen, H. S. & Hünenberger, P. H. Using the local elevation method to construct optimized umbrella sampling potentials: Calculation of the relative free energies and interconversion barriers of glucopyranose ring conformers in water. *J Comput Chem* **31**, 1–23 (2010).

35. Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H. & Kollman, P. A. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J Comput Chem* **13**, 1011–1021 (1992).

36. Souaille, M. & Roux, B. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Comput Phys Commun* **135**, 40–57 (2001).

37. Bohner, M. U. & Kästner, J. An algorithm to find minimum free-energy paths using umbrella integration. *J Chem Phys* **137**, 034105 (2012).

38. Kästner, J. Umbrella integration with higher-order correction terms. *J Chem Phys* **136**, 234102 (2012).

39. Jiang, W., Luo, Y., Maragliano, L. & Roux, B. Calculation of free energy landscape in multi-dimensions with hamiltonian-exchange umbrella sampling on petascale supercomputer. *J Chem Theory Comput* **8**, 4672–4680 (2012).

40. Higo, J., Nakajima, N., Shirai, H., Kidera, A. & Nakamura, H. Two-Component

Multicanonical Monte Carlo Method for Effective Conformation Sampling. *J Comput Chem* **18**, (1997).

41. Sugita, Y., Kitao, A. & Okamoto, Y. Multidimensional replica-exchange method for free-energy calculations. *J Chem Phys* **113**, 6042 (2000).

42. Bartels, C. & Karplus, M. Multidimensional adaptive umbrella sampling: Applications to main chain and side chain peptide conformations. *J Comput Chem* **18**, 1450–1462 (1997).

43. Hayami, T., Kasahara, K., Nakamura, H. & Higo, J. Molecular dynamics coupled with a virtual system for effective conformational sampling. *J Comput Chem* **39**, 1291–1299 (2018).

44. Higo, J., Kamiya, N., Sugihara, T., Yonezawa, Y. & Nakamura, H. Verifying trivial parallelization of multicanonical molecular dynamics for conformational sampling of a polypeptide in explicit water. *Chem Phys Lett* **473**, 326–329 (2009).

45. Kasahara, K. *et al.* myPresto/omegagene: a GPU-accelerated molecular dynamics simulator tailored for enhanced conformational sampling methods with a non-Ewald electrostatic scheme. *Biophys Physicobiol* **13**, 209–216 (2016).

46. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* **23**, 327–341 (1977).

47. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Dinola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J Chem Phys* **81**, 3684 (1998).

48. Fukuda, I., Yonezawa, Y. & Nakamura, H. Molecular dynamics scheme for precise estimation of electrostatic interaction via zero-dipole summation principle. *J Chem Phys* **134**, 164107 (2011).

49. Fukuda, I., Kamiya, N., Yonezawa, Y. & Nakamura, H. Simple and accurate scheme to compute electrostatic interaction: Zero-dipole summation technique for molecular system and application to bulk water. *J Chem Phys* **137**, 054314 (2012).

50. Kamiya, N., Fukuda, I. & Nakamura, H. Application of zero-dipole summation method to molecular dynamics simulations of a membrane protein system. *Chem Phys Lett* **568–569**, 26–32 (2013).

51. Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics* **65**, 712–725 (2006).

52. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem*

Phys **79**, 926 (1998).

53. Schoenherr, C. J. & Anderson, D. J. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science (1979)* **267**, 1360–1363 (1995).
54. Chong, J. A. *et al.* REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell* **80**, 949–957 (1995).
55. Bruce, A. W. *et al.* Genome-wide analysis of repressor element 1 silencing transcription factor/neuron-restrictive silencing factor (REST/NRSF) target genes. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 10458–10463 (2004).
56. Rockowitz, S. & Zheng, D. Significant expansion of the REST/NRSF cistrome in human versus mouse embryonic stem cells: potential implications for neural development. *Nucl. Acids Res.* **43**, 5730–5743 (2015).
57. Lawinger, P. *et al.* The neuronal repressor REST/NRSF is an essential regulator in medulloblastoma cells. *Nat. Med.* **6**, 826–831 (2000).
58. Fuller, G. N. *et al.* Many human medulloblastoma tumors overexpress repressor element-1 silencing transcription (REST)/neuronrestrictive silencer factor, which can be functionally countered by REST-VP16. *Mol. Cancer Ther.* **4**, 343–349 (2005).
59. Dhall, G. Medulloblastoma. *J. Child Neurol.* **24**, 1418–1430 (2009).
60. Conti, L. *et al.* REST controls self-renewal and tumorigenic competence of human glioblastoma cells. *PLoS One* **7**, e38486 (2012).
61. Kamal, M. M. *et al.* REST regulates oncogenic properties of glioblastoma stem cells. *Stem Cells* **30**, 405–414 (2012).
62. Zuccato, C. *et al.* Huntingtin interacts with REST/NRSF to modulate the transcription of NRSE-controlled neuronal genes. *Nat. Genet.* **35**, 76–83 (2003).
63. Zuccato, C. & Cattaneo, E. Role of brain-derived neurotrophic factor in Huntington's disease. *Prog. Neurobiol.* **81**, 294–330 (2007).
64. Bithell, A., Johnson, R. & Buckley, N. J. Transcriptional dysregulation of coding and non-coding genes in gellular models of Huntington's disease. *Biochem. Soc. Trans.* **37**, 1270–1275 (2009).
65. Buckley, N. J., Johnson, R., Zuccato, C., Bithell, A. & Cattaneo, E. The role of REST in transcriptional and epigenetic dysregulation in Huntington's disease. *Neurobiol. Dis.* **39**, 28–39 (2010).
66. Uchida, H., Ma, L. & Ueda, H. Epigenetic gene silencing underlies C-fiber dysfunctions in neuropathic pain. *J. Neurosci.* **30**, 4806–4814 (2010).
67. Willis, D. E., Wang, M., Brown, E., Fones, L. & Cave, J. W. Selective repression of gene expression in neuropathic pain by the neuronrestrictive silencing

factor/repressor element-1 silencing transcription (NRSF/REST). *Neurosci. Lett.* **625**, 20–25 (2016).

68. Suo, H. *et al.* NRSF is an essential mediator for the neuroprotection of trichostatin A in the MPTP mouse model of Parkinson's disease. *Neuropharmacology* **99**, 67–78 (2015).

69. Katayama, Y. *et al.* CHD8 haploinsufficiency results in autistic-like phenotypes in mice. *Nature* **537**, 675–679 (2016).

70. Ueda, H. *et al.* A mimetic of the mSin3-binding helix of NRSF/REST ameliorates abnormal pain behavior in chronic pain models. *Bioorg. Med. Chem. Lett.* **27**, 4705–4709 (2017).

71. Ooi, L. & Wood, I. C. Chromatin crosstalk in development and disease: lessons from REST. *Nat. Rev. Genet.* **8**, 544–554 (2007).

72. Naruse, Y., Aoki, T., Kojima, T. & Mori, N. Neural restrictive silencer factor recruits mSin3 and histone deacetylase complex to repress neuron-specific target genes. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 13691–13696 (1999).

73. Bansal, N., David, G., Farias, E. & Waxman, S. Emerging roles of epigenetic regulator Sin3 in cancer. *Adv. Cancer Res.* **130**, 113–135 (2016).

74. Nomura, M., Uda-Tochio, H., Murai, K., Mori, N. & Nishimura, Y. The neural repressor NRSF/REST binds the PAH1 domain of the Sin3 corepressor by using its distinct short hydrophobic helix. *J. Mol. Biol.* **354**, 903–915 (2005).

75. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331 (1999).

76. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208 (2005).

77. Sugase, K., Dyson, H. J. & Wright, P. E. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature* **447**, 1021–1025 (2007).

78. Leone, S. *et al.* SAR and QSAR study on 2-aminothiazole derivatives, modulators of transcriptional repression in Huntington's disease. *Bioorg. Med. Chem.* **16**, 5695–5703 (2008).

79. Charbord, J. *et al.* High throughput screening for inhibitors of REST in neural derivatives of human embryonic stem cells reveals a chemical compound that promotes expression of neuronal genes. *Stem Cells* **31**, 1816–1828 (2013).

80. Conforti, P. *et al.* Binding of the repressor complex REST-mSIN3b by small molecules restores neuronal gene tanscription in Huntington's disease models. *J. Neurochem.* **127**, 22–35 (2013).

81. Kurita, J., Hirao, Y., Miyata, N. & Nishimura, Y. NMR Screening of mSin3B

Binding Compounds for the Interaction Inhibition with a Neural Repressor, NRSF/REST. *Modern Magnetic Resonance* 1–22 (2017) doi:10.1007/978-3-319-28275-6_64-1.

82. Kurita, J. ichi, Hirao, Y., Nakano, H., Fukunishi, Y. & Nishimura, Y. Sertraline, chlorprothixene, and chlorpromazine characteristically interact with the REST-binding site of the corepressor mSin3, showing medulloblastoma cell growth inhibitory activities. *Sci. Rep.* **8**, 13763 (2018).

83. Hayami, T., Higo, J., Nakamura, H. & Kasahara, K. Multidimensional virtual-system coupled canonical molecular dynamics to compute free-energy landscapes of peptide multimer assembly. *J. Comput. Chem.* **40**, 2453–2463 (2019).

84. Higo, J. *et al.* GA-guided mD-VcMD: a genetic-algorithm-guided method for multi-dimensional virtual-system coupled molecular dynamics. *Biophys. Physicobiol.* **17**, 161–176 (2020).

85. Higo, J. *et al.* Molecular interaction mechanism of a 14-3-3 protein with a phosphorylated peptide elucidated by enhanced conformational sampling. *J. Chem. Inf. Model.* **60**, 4867–4880 (2020).

86. Mashimo, T. *et al.* Molecular dynamics simulations accelerated by GPU for biological macromolecules with a non-Ewald scheme for electrostatic interactions. *J. Chem. Theory Comput.* **9**, 5599–5609 (2013).

87. Ikebe, J. *et al.* Theory for trivial trajectory parallelization of multicanonical molecular dynamics and application to a polypeptide in water. *J. Comput. Chem.* **32**, 1286–1297 (2011).

88. Morishita, T. Fluctuation formulas in molecular-dynamics simulations with the weak coupling heat bath. *J. Chem. Phys.* **113**, 2976–2982 (2000).

89. Kamiya, N., Watanabe, Y. S., Ono, S. & Higo, J. AMBER-based hybrid force field for conformational sampling of polypeptides. *Chem Phys Lett* **401**, 312–317 (2005).

90. Joung, I. S. & Cheatham, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **112**, 9020–9041 (2008).

91. Frisch, M. J. *et al.* Gaussian09 Revision D.01, Gaussian Inc. Wallingford CT. *Gaussian 09 Revision C.01* Preprint at (2010).

92. Bayly, C. I., Cieplak, P., Cornell, W. D. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **97**, 10269–10280 (1993).

93. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general Amber force field. *J Comput Chem* **25**, 1157–1174 (2004).

94. Bayly, C. I. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197 (1995).

95. Kollman, P. *et al.* The development/application of a 'minimalist' organic/biochemical molecular mechanic force field using a combination of ab initio calculations and experimental data. *Computer Simulation of Biomolecular Systems* 83–96 (1997) doi:10.1007/978-94-017-1120-3_2.

96. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* **125**, 1731–1737 (2003).

97. Bekker, G. J., Kawabata, T. & Kurisu, G. The biological structure model archive (BSM-Arc): an archive for in silico models and simulations. *Biophys Rev.* **12**, 371–375 (2020).

Acknowledgement

First of all, I would like to express my gratitude to Prof. Haruki Nakamura for giving me research guidance as a supervisor from the beginning of my research life at Osaka University. I would also like to thank Prof. Atsushi Nakagawa for taking over my supervisor and for his support until I completed my degree.

I would like to thank Prof. Yoshifumi Nishimura and Dr. Jun-ichi Kurita for providing the experimental results and giving advice regarding the research of mSin3B. I would also like to thank Dr. Yoshifumi Fukunishi for his cooperation in the selection of ligand molecules based on the results of *in silico* screening for the research of mSin3B. I'd also want to thank Dr. Narutoshi Kamiya for his cooperation in creating the force fields of the ligand molecules.

I would like to thank Dr. Takeshi Kawabata for providing me with various consultations and advice, not limited to research.

I express my deep gratitude to Dr. Jun-ichi Higo and Dr. Kota Kasahara for their excellent assistance and direction in conducting this study.

Publication list

1) Hayami, T., Kasahara, K., Nakamura, H., Higo, J., "Molecular dynamics coupled with a virtual system for effective conformational sampling", *Journal of Computational Chemistry*, (2018), 1291-1299, 39(19)

2) Hayami, T., Higo, J., Nakamura, H., Kasahara, K., "Multidimensional virtual-system coupled canonical molecular dynamics to compute free-energy landscapes of peptide multimer assembly", *Journal of Computational Chemistry*, (2019), 2453-2463, 40(28)

3) Hayami, T., Kamiya, N., Kasahara, K., Kawabata, T., Kurita, J., Fukunishi, Y., Nishimura, Y., Nakamura, H., Higo, J., "Difference of binding modes among three ligands to a receptor mSin3B corresponding to their inhibitory activities", *Scientific Reports*, (2021), 11(1)