博士論文

Doctoral Thesis

Association and dissociation simulations of bio-molecular complex using parallel cascade selection molecular dynamics

(並列カスケード選択分子動力学法による生体分子の会合・解離シミュレーション)

東京大学大学院新領域創成科学研究科

メディカル情報生命専攻

チャン　フ　ズイ

In memory of my father

To my mother and my sister
With eternal love and appreciation

Abstract

Sampling conformations of protein complexes during association and dissociation processes is a crucial step to estimate the binding free energy and other kinetic properties from association/dissociation pathways. This is a challenging problem for the classical Molecular Dynamics (MD) simulation because the time scale of these processes exceeds the limit of current computation. Therefore, enhanced sampling techniques play an important role to generate sufficient data for the free energy analysis. For example, Steered Molecular Dynamics (SMD) with Umbrella Sampling (US) [Ramirez et al., Methods Enzymol. (2016)], Replica Exchange Umbrella Sampling (REUS) [Sugita et al., J. Chem. Phys (2000)], Targeted MD (TMD) [Schlitter et al., J. Mol. Graph. (1994)], Parallel Cascade Selection Molecular Dynamics (PaCS-MD) [Harada and Kitao, J. Chem Phys. (2011)] and other methods not listed here are used for this purpose. Recently, Yamashita and Fujitani showed that protein structures were distorted when dissociation of lysozyme (enzyme) and HyHEL-10 (inhibitor) was simulated by SMD using a steering force applied to the center of mass (COM) of the protein, which led overestimation of the potential of mean force (PMF) with the following US. This can be considered as the artifact caused by SMD. In contrast to SMD, PaCS-MD performs conformational sampling by cycles of distinct multiple Molecular Dynamics (MD) simulations without applying any bias force to the system. It enhances the sampling by selecting the MD snapshots closest to the destination state and by restarting the MD simulations from the selected snapshots with the velocity re-randomization. PaCS-MD was shown to be very successful in efficient sampling of protein domain motions. Here in this thesis, we describe unbiased association and dissociation simulations by PaCS-MD.

We first show that PaCS-MD dissociated a small ligand, tri-N-acetyl-D-glucosamine (triNAG), from hen egg white lysozyme (LYZ) very efficiently. We performed PaCS-MD trials with 3 different simulation settings: PaCS-MD$^{10,0.1}$ (ten 0.1 ns MDs per cycle), PaCS-MD$^{100,0.1}$ (hundred 0.1 ns MDs) and PaCS-MD^{10,1} (ten 1.0 ns MDs). We found that PaCS-MD is 5 times faster than SMD. In combination with Markov State Model (MSM), we calculated the binding free energy directly from the

PaCS-MD trajectories. In comparison, binding free energy was also calculated by the analysis of SMD trajectories using the Jarzynski equality [Jarzynski., Phys. Rev. Lett. (1997)]. Although SMD/Jarzynski overestimated the binding free energy, PaCS-MD/MSM yielded the results in good agreement with experimental results. We also examined the effects of the number of replicas, the length of each MD, the velocity re-randomization, and the selection of snapshots on PaCS-MD sampling. We found that the increase of the number of replicas reduced the number of cycles required for dissociation because the probability of observing rare events is proportional to the number of replicas. The velocity re-randomization enhances the sampling in the bound state as it acts as a perturbation to raise the occurrence of rare events (dissociation).

We next applied PaCS-MD to the dissociation of MDM2 protein and trans-activation domain of p53 (TAD-p53). Binding free energy of MDM2/TAD-p53 calculated by PaCS-MD/MSM was 40.5 ± 1.7 kJ/mol, which almost agrees with experimental value 37.7 ± 1.7. kcal/mol. Our result is more accurate than the value calculated by the MMGBSA method, 68.2 kJ/mol [Dastidar et al., JACS (2008)]. We found the calculated binding free energy for each trial is strongly dependent on the dissociation pathway of TAD-p53, which is related to the dissociation of the key residues PHE19 and TRP23 of TAD-p53 involved in π-π stacking interactions between TAD-p53 and MDM2.

We also employed PaCS-MD for simulating association and dissociation process of MDM2/TAD-p53, which can be considered as a flexible-body docking simulation. We used the switching condition between the dissociation and association simulations as follows: if the association simulation does not make any progress for continuous 20 ps, it will switch to the dissociation simulation. When the inter COM distance between MDM2 and TAD-p53 reaches 2.0 nm longer than the last switching point, the association simulation will start. We performed 274 cycles of PaCS-MD and examined whether generated structures of TAD-p53 and MDM2 complex are similar to the crystal complex structure and found that the minimum RMSD was 0.429 nm. In addition, TAD-p53 could bind to the correct binding interface without the guiding force. We further examined 4 representative structures selected from all the bound conformations. Although the two key π-π stacking interactions were not formed in these structures,

residual contacts are in agreement with those in the crystal structure with the binding interface RMSD of 0.243 nm. To predict the bound conformation without prior-knowledge of the crystal structure, we examined if the conformation similar to the correct bound conformation can be identified as the lowest free energy structure. We built MSM based on the trajectories of distance RMSD (dRMSD) from the initial conformation of MDM2/TAD-p53 in the unbound state and calculated the Potential of Mean Force (PMF). We found that dRMSD of the lowest PMF position was 4.21 nm, which the corresponding structure was identical to the structure with the lowest interface RMSD from the crystal structure. Therefore, we can select the best structure based on the calculated RMSD.

In conclusion, PaCS-MD algorithm was shown to be an efficient unbiased enhanced sampling tool which can be applied to bio-molecular complexes and is highly suitable for distributed computing. Overall, PaCS-MD is faster in computational time than the other biased sampling techniques. We are currently making an effort to apply PaCS-MD for reducing total simulation time of flexible-body docking simulation.

List of publications

1) Duy Phuoc Tran, Kazuhiro Takemura, Kazuo Kuwata, Akio Kitao, “Protein-Ligand Dissociation Simulated by Parallel Cascade Selection Molecular Dynamics”, Journal of Chemical Theory and Computation 14, 404 (2018)

List of abbreviations

COM: Center of mass

MD: Molecular dynamics

NPT ensemble: isothermal-isobaric ensemble

NVT ensemble: canonical ensemble

PaCS-MD: Parallel Cascade Selection Molecular Dynamics

PaCS-MD^{x,y}: Parallel Cascade Selection Molecular Dynamics with x replicas and its simulation time for each replica is y ns

MSM: Markov State Model

SMD: Steered Molecular Dynamics

LYZ: Hen-egg white lysozyme

triNAG: tri-*N*-acetyl-D-glucosamine

PDB: Protein data bank

PMF: Potential of mean force

MM/PB-SA: Molecular mechanics/Poisson Boltzmann –Surface Area

WHAM: Weighted Histogram Analysis Method

RMSD: Root Mean-Squared Deviation

dRMSD: distance Root Mean-Squared Deviation

MFPT: Mean First Passage Time

PMF: Potential of Mean Force

Table of Contents

Chapter 1. Introduction

1. The importance of understanding association and dissociation events of biomolecular complex

Association and dissociation of bio-molecular complexes play an important role in biological phenomena. For example, G Protein Coupled Receptor (GPCR) family, a seven transmembrane helices receptor, exists to be the communication channel between intra-cellular and extra-cellular environment. Upon binding of a ligands in the case of Adenosine A2A receptor, GPCRs changes to be in active-intermediate state, and later in a fully active state upon the association of G Proteins[1]. After being in fully active state, following cascade events will take place that makes the organism to adapt with the external signaling[2]. Specifically, when drinking coffee, caffeine ligands bind to Adenosine A2A receptor, a subtype of GPCR, and deactivate Adenosine A2A leading to the reduction of stress response[3,4]. As shown in this example, understanding the association and dissociation of bio-molecular complex is the crucial works for thoroughly understanding the given biological phenomena.

It is obvious that there is a need for determining quantities to describe the strength of the binding in energy unit, e.g., "binding free energy". For instance, let's consider two biomolecules A and B that can bind to each other via a reaction as following:

$$A + B \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} AB \tag{1}$$

Here we denote the so-called quantity the association rate constant k_{on} to describe the rate of binding of AB, and the dissociation rate constant k_{off} to represent the separation rate of the two molecules A and B from their complex AB with the concentrations [A], [B], [AB] respectively. The reaction in (1) is in equilibrium only if the concentrations [AB] does not change in the vicinity of time as follow:

$$\frac{d[AB]}{dt} = [A].[B].k_{on} - [AB].k_{off} = 0 \tag{2}$$

Here, one can define the equilibrium association constant K_a or equilibrium dissociation constant K_d as the fraction of k_{on} and k_{off}.

$$K_d = \frac{1}{K_a} = \frac{^k{}_{off}}{k_{on}} = \frac{[A].[B]}{[AB]} \quad (3)$$

From equation (3), one can directly convert the equilibrium constant to the free energy difference via the following equation.

$$\Delta G = -\frac{1}{\beta} ln(C^0 K_b) = -\frac{1}{\beta} ln(C^0/K_d) \quad (4)$$

in which β is the thermodynamic temperature and C^0 is the standard concentration which is equal to 1 M.

Equations (3) and (4) show the relation between the concentrations of subtances in the samples and the binding free energy of the given complex. In experiment, the common method to determine the binding free energy is from the estimation of equilibrium constant such as Isothermal Titration Calorimetry (ITC)[5], Surface Plasmon Resonance (SPR)[6], flourescene quenching method[7], and binding assay[8,9]. ITC experiment is the only and the most common method directly measuring the binding kinetics of a given bio-molecular complex[5]. ITC experiment measures the energy consumption to maintain the temperature in the sample cell and adiabatically identical reference cell while increasing the ligand concentration in sample cell. Although ITC is considered as high pricision method, the amount of sample used is high that limits the applicability of the methods to the protein complexes which is difficult to express massively. Moreover, the experimental methods to determine the binding free energy in general that cannot be done extensively due to the difficulty in experiment setup procedure. Therefore, computational methods for binding free energy computation are generally essential for the initial stage of the research i.e. bio-molecular interaction design in general speaking or computational drug design specifically. In addition, computational methods can provide additional information on structural and dynamic properties of the given biomolecular system, which requires enormous efforts in crystallography. Up to now, the extensive development of either computing resource, accuracy of calculation methods or parameters allow *in silico* experiment to reduce total budget for research in screening and structural optimization.

2. Classification of binding free energy calculation and limitation of current methods

Binding free energy calculation which has a long history of methodological development can be classified into four catalogues: thermodynamic integration, sampling based method, non-equilibrium dynamics and adaptive biasing technique [10]. Thermodynamic integration is a widely-used method which takes an advantage of the adiabatic evolution of statistical average along reaction coordinates to calculate the binding free energy[11]. It can be defined as the total change of free energy from the unbound state A to the bound state B as in the following equation:

$$\Delta G(A \to B) = \int_0^1 \frac{\partial G(\lambda)}{\partial \lambda} d\lambda, \quad (5)$$

in which λ is the coupling parameter of the given system. The change in λ adiabatically leads the system from state A to state B. In contrast, Zwanzig proposed the alchemical free energy perturbation which decomposes the free energy change into multiple intermediate steps[12]. Later free energy perturbation method was extended to reaction coordinate based methods including Umbrella Sampling[13]. The non-equilibrium dynamics method is generally based on the Jarzynski equality[14]. It describes the relation between the applied work to the system and its free energy change. Adaptive biasing dynamics monitors the reaction coordinates and prevents trap around free energy minima already explored by using the dynamics forces as in metadynamics[15] and Wang-Landau methods[16]. Overall, success of the above methods to accurately calculate binding free energy calculation[17] depends on a) whether the chosen model Hamiltonian is suitable, b) whether the sampling is correct and sufficient, and c) whether the estimator for the free energy difference is appropriate. Basically, the consideration a) in term of the force fields has been much improved in last decades including AMBER[18], GROMOS[19], CHARMM[20], and OPLS[21] force fields. They yield reliable results which agree with experiments. Free energy calculations can be conducted by combinations of Steered Molecular Dynamics simulation with Umbrella Sampling (SMD/US) [22], MD with restraining potentials[23–26], replica exchange umbrella sampling (REUS)[27], and targeted MD (TMD)[28] with US (TMD/US)[29] and more not listed here. From the point of view for sampling classification, these methods are mostly based on the MD

simulation using biased force to accelerate sampling of dissociation or association in comparison to the non-biased MD. However, sampling of association process is difficult for free energy computation due to trapping of local minima of free energy landscape and the complexity of the binding pathways. Therefore, simulation of dissociation process is more suitable to estimate the binding free energy.

The most popular approach by biased-MD-based methods is to incorporate SMD with US. In SMD/US, bias forces are used to generate ligand dissociation pathways in SMD by pulling the ligand through an artificial harmonic spring connecting the ligand and a particle which moves with constant velocity. After that, US explores the overlapped local conformational spaces which are sampled with multiple windows along the generated pathway by the SMD. The binding free energy is then obtained as the potential of mean force between the bound and unbound states. However, by applying the biased force to the system, the protein structures are distorted and stay in metastable states that cannot be recovered in US calculation[29]. In the case of lysozyme and HyHEL-10 complex, large structural distortion was avoided by using multi-step TMD[29]. Consequently, the dissociation pathway generated by SMD contains artifacts, especially for large or very flexible biomolecules with high degrees of freedom, which leads to the system unescapable from the metastable state leading to higher estimation of binding free energy. As a result, there is a need of alternative methods for binding free energy calculation to mimic the above problem.

3. Free energy calculation without using any biased force

Here we proposed to examined following features to improve the binding free energy computation. First, to overcome the above problems of sampling, we expected to obtain more natural pathways by unbiased simulation, which facilitate the estimation of binding free energy closer to the correct value than those obtained by the other scheme. In fact, association and dissociation of bio-molecular complex occurs as a rare event that spans up to second timescale[30]. Consider using CPU with dual core at 2.6 GHz, a typical time to reach to second timescale for capturing these rare event would generally take upper 1.4 million years[30]. Therefore, an enhanced sampling technique is prerequisite to observe these rare event in computer simulation. For summary, there is a need of an unbiased enhanced sampling technique to simulate the dissociation and association of biomolecular complex.

Recently, the Parallel Cascade Selection Molecular Dynamics (PaCS-MD) simulation was introduced as the method that satisfis these needs for unbiased sampling. PaCS-MD was first introduced in 2013 and was applied successfully to folding of chignolin protein and conformational transition of T4 lysozyme which captured rare events[31]. Folding time of chignolin protein was found to be 0.4 μs or slower in classical MD simulation whilst that occured within 2 ns of PaCS-MD. Later, alternative version of PaCS-MD without prior knowledge of the target conformation i.e nontargeted PaCS-MD (nt-PaCS-MD) was introduced[32]. In nt-PaCS-MD, the Gram-Schmidt orthogonalization is applied to select the significant conformations within the cycle. nt-PaCS-MD was successful in obtaining the native state of chignolin, sampling the open-closed transition of T4 lysozyme within nanosecond timescale. In addition, the so-called PaCS-Fit method, a derivative version of PaCS-MD, has successfully fit small-angle X-ray scattering and electron microscopy data[33].

Selection in PaCS-MD is a key for acceleration in sampling which enhances the probability of transitions between microstates. The transition probability between microstates is very useful in building a transition matrix, a part of Markov State Model (MSM). From MSM, we can construct a kinetic model of our given system. Moreover, one can directly extract the equilibrium free energy difference via stationary

eigenvector of the transition matrix from MSM. Therefore, we hypothesized that we can calculate the binding free energy in agreement with experimental data, directly from the PaCS-MD trajectories, without additional sampling such as US. The total simulation time for obtaining binding free energy accordingly decreases compared to SMD/US.

To examine this, we first carried out the protein/ligand dissociation simulation using PaCS-MD. We chose Lysozyme/triNAG complex to be the target due to the availability of preceeding computational and experimental results and the suitable system size as the first test case. Then, we extended the method to more difficult case, protein/peptide dissociation, which is MDM2 protein in complex with transactivation domain of p53. Next, we applied the PaCS-MD scheme to flexible protein-ligand docking.

Chapter 2. Simulation methods for binding free energy calculation

1. Parallel cascade selection molecular dynamics simulation method

PaCS-MD consists of cycles in which multiple independent parallel simulations are conducted, starting from selected initial configurations of given system without applying any bias to the system[31]. The simulation first starts with a single long MD simulation to generate inputs for the parallel simulations later. Here, n_{rep} is defined as the number of replicas of parallel simulations and t_{cyc} is the length of each simulation in each cycle. The selected snapshots of each cycle with any pre-defined selection criterion are then to be employed as starting structures for the next cycle, which is started with randomized initial velocities to obey the Maxwell-Boltzmann distribution. The procedure is repeated until the generated snapshots approach to a target. To generate dissociation pathways, we only used inter-center of mass (inter-COM) distance for the selection. In addtion, we also included the initial snapshots in ranking. Reactive trajectories are defined as the trajectories which connect the initial bound state and the final unbound state along dissociation pathways concatenated fragments of the selected MD trajectories[31]. An example of PaCS-MD is shown in Fig 1. After rank-ordering inter-COM distance in descending order, top n_{rep} snapshots are selected as the input coordinates for the MD simulation of the next cycle (the first table in the left-hand side of Fig 1). All of the generated snapshots in the next cycle are then rank-ordered and selected as shown in the center table of Fig 1. The yellow highlights in Fig 1 show the survived snapshots that plays the role of links between cycles. One can see that not all the selected snapshots survived after a few cycles. In this thesis, we sample snapshots every 0.5 ps from the generated trajectories.

Although proven to be a highly efficient unbiased enhanced sampling technique, the mechanism of acceleration in PaCS-MD has not been thoroughly examined yet. Moreover, how the t_{cyc} and n_{rep} affect the sampling effiency is still an open question in PaCS-MD. Generated trajectories in PaCS-MD are continuous in conformational space, however, dynamics in the reactive trajectories each short MD simulations in each cycle have not been examined yet. We will discuss these questions in detail in chapter 3 of this thesis.

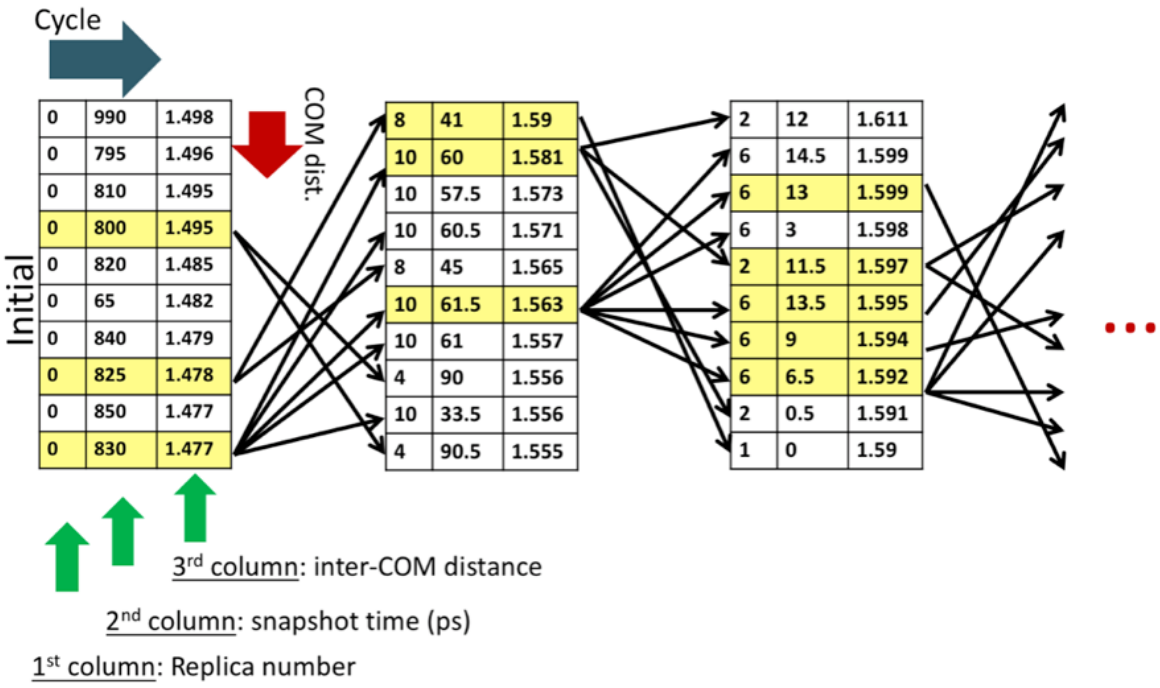


Fig 1. Illustration of Parallel Cascade Selection Molecular Dynamics simulation. Each table represents replica number (the first column), snapshot number (the second column), and inter-COM distance (the third column), in each cycle in PaCS-MD. The table only shows the selected snapshot after the ranking in each cycle. Yellow highlight shows the survived snapshot in PaCS-MD.

2. Markov State Model in combination with PaCS-MD as a state-of-art free energy estimation tool

The MSM is a discrete-state stochastic kinetic model of the observed process and is a powerful tool to obtain insights for linking experimental and simulation data[34,35]. MSM solves the Master equation, in which kinetics is described by the rates of transition among N discrete states [36]. The states here can be thousands to millions but should not be limited to a few states. Generally, there are four steps to build the MSM from MD trajectories: preparing dataset, building microstates, building the transition matrix, and validating the generated MSM.

Datasets for building MSM can be obtained from MD simulation. To make use the computing resource and availability of memory, one may need to map the higher dimensional data generated as MD trajectories to lower dimensional space by principle component analysis (PCA)[37], time-lagged independent component analysis (TICA)[38–40] or either the coarse-grained model. Next step is to assign the microstates. The processed dataset is then clustered into the microstates which provide transition rates between them in a kinetically meaningful manner[36]. In this thesis, we applied the k-means clustering[41] to the inter-COM distance for determininig the microstates. The k-means clustering is a very fast clustering method using Lloyd's algorithm[42]. It consists of four steps: first the cluster centroids are assigned randomly; second the distance between each datapoint and cluster centers (centroids) are calculated; third each datapoint is assigned to a cluster with the nearest centroids; fourth the new centroids are calculated and the procedure from second step to fourth step is repeated until convergence of centroids is achieved.

Consider a system having a set of microstates $\{S_i\}$ and a transition from microstate S_i to microstate S_j is observed. After obtaining the microstates by using clustering, the transition matrix between the microstates

$$T = \{T_{ij}\} = \{P[x(t+\tau) \in S_j | x(t) \in S_i]\} \quad (6)$$

was estimated. Each matrix element T_{ij} is calculated between a pair of microstates (i and j) with a predetermined lag time τ, using the maximum likelihood estimation

procedure described in the reference [43] by maximizing the likelihood probability $\mathcal{L}(T)$ as followed.

$$\mathcal{L}(T) = P(S|T) = \prod_{t=0}^{N-\tau} T_{S_t,S_{t+\tau}} = \prod_{i,j}^{k} T_{ij}^{C_{ij}} \quad (7)$$

In equation (6), $x(t)$ and $x(t+\tau)$ represent the coordinates at time t and $(t+\tau)$, respectively. To build a good MSM, the lagtime τ should be chosen with carefulness. The stationary probabilities of the microstates can be calculated as the eigenvector $p = \{p_i\}$ of T. The equilibrium free energy of microstate i can be obtained as $-(\ln p_i)/\beta$ [44].

To build each MSM, we employed all the full MD trajectories generated by each trial of PaCS-MD, regardless of whether or not the snapshots of the trajectories are selected for the next cycle. It should be noted that the selected and non-selected trajectories together provide significant information to estimate transition probabilities between microstates. The obtained probabilities as a function of the inter-COM distance are averaged and shown in the result. We used MSMBuilder 3 for MSM[45].

3. Steered molecular dynamics in combination with Jarzynski equality

SMD is an fast and direct methods to pull a ligand out of its binding pocket to obtain the dissociation pathway[46]. In SMD, a spring is used to stick one head to the center of mass on the ligand while the other head is sticked to the dummy atom that moves with a given constant velocity as shown in Fig 2.

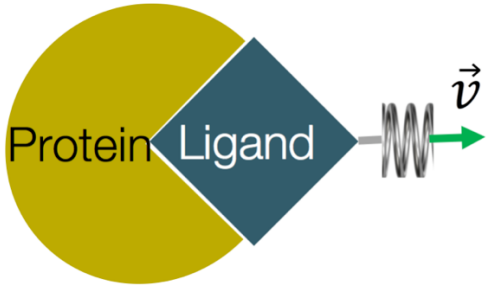


Fig 2. Velocity constant steered molecular dynamics simulation for pulling a ligand out of its complex with protein

The effect of the spring attached to the ligand is described by a harmonic potential as in equation (8).

$$\frac{1}{2}k[\xi(t) - \lambda(t)]^2 \quad (8)$$

where k is the force constant of the spring, $\xi(t)$ is the inter-COM distance between LYZ and triNAG at SMD time t, and $\lambda(t) = \lambda_0 + vt$ is the distance between the COM of LYZ and the dummy atom (λ_0 is the initial distance) and v is the pulling velocity, respectively. In SMD, the ligand is pulled out from the binding site as in Atomic Force Microscope (AFM) experiment. Two deterministic parameters that lead to the success of SMD are the velocity of dummy atom or so-called pulling speed and the force constant of the spring. One may consider that the pulling speed can be the same as in AFM; However, it is impossible to reproduce experimental pulling speed because simulation time is a few orders shorter than the time spent in AFM.

From SMD simulation of dissociation, one can directly estimate the binding free energy by simple but effective relation, the Jarzynski equality, as in equation (9)[14].

$$e^{-\beta\Delta G} = \langle e^{-\beta W} \rangle \quad (9)$$

$$W_{0\to t} = -kv\int_0^t dt'[\xi(t') - \lambda(t')] \quad (10)$$

in which $\langle ... \rangle$ indicates the statistical average of the quantities. The simple relation in equation (9) implies that we can directly calculate the equilibrium information (binding free energy ΔG) from the ensemble of non-equilibrium quantity (work acts on the system) that can be calculated from equation (10).

4. Weighted Histogram Analysis Method with Umbrella Sampling

WHAM and US can be considered as an extension of free energy perturbation method[13]. In US, the Hamiltonian of a given system H_0 and a function H_λ is added with a coupling parameter λ and a modified potential V_i in equation (11).

$$H_\lambda(x) = \sum_{i=0}^{L} \lambda_i V_i(x) \tag{11}$$

where x is atomic coordinates and $\lambda_0 = 1$. The unbiased system with $\lambda_i = 0$ is identical to the H_0, and Therefore, the probability density $P_{\{\lambda\}}(\xi)$ due to the reaction coordinate ξ can be computed based on the simulation with Halmitonian H_λ[13] is:

$$P_{\{\lambda\}}(\xi) = e^{-\beta W_{\{\lambda\}}(\xi)} = \langle \delta[\xi - \xi(x)] \rangle_{\{\lambda\}} \tag{12}$$

Then, the unbiased probability of the system where $\lambda = 0$ can be calculated via

$$P_{\{0\}}(\xi) = \frac{Z_{\{\lambda\}}}{Z_{\{0\}}} \langle \delta[\xi - \xi(x)] \prod_{i=1}^{L} e^{\beta \lambda_i V_i(x)} \rangle_{\{\lambda\}} \tag{13}$$

where Z is the partition function and $V_i(x)$ is the restrained potential of the atomic coordinates with respect to the reaction coordinates. Based on Weighted Histogram Analysis Method, Kumar *et al.* derived the evolution of free energy with reaction coordinate via R simulations having n_i of simulation i at the temperature $T_i = 1/k_B\beta_i$, where k_B is the Boltzmann constant[13].

$$\Delta G_i = -ln\left(\sum_{k=1}^{R} \sum_{t=1}^{n_k} \frac{exp\left(-\beta_i \sum_{j=0}^{L} \lambda_{j,i} V_{j,t}^{(k)}\right)}{\sum_{m=1}^{R} n_m exp\left[f_m - \beta_m \sum_{j=0}^{L} \lambda_{j,m} V_{j,t}^{(k)}\right]} \right) \tag{14}$$

The procedure for calculating binding free energy using WHAM US can be summarized here. US calculates free energy from a probability distribution in equilibrium. Restrained MD simulations with the umbrella potential V are conducted around different points along a reaction coordinate, here is inter-COM distance. In this work, V was applied to the inter-COM distance d between protein and ligand along the dissociation pathway. The free energy profile $\Delta G(d)$ can be calculated by WHAM[13].

Chapter 3. Dissociation of small ligand from its complex with protein

1. Introduction

In this work, we employed PaCS-MD[31,47] to generate ligand dissociation pathways without applying force biases. We demonstrate that PaCS-MD can be used to simulate protein-ligand dissociation within tens of nanoseconds by employing a longer inter-molecular distance as the target quantity for the selection of the initial structures without applying force bias. The dissociation pathways generated by PaCS-MD are comparable to those of SMD. The free energy change along the dissociation pathways is calculated by all trajectories obtained by PaCS-MD in combination with the Markov state model (MSM). For comparison, alternative combinations for free energy calculation are also employed such as PaCS-MD with US (denoted as PaCS-MD/US), SMD and US (denoted as SMD/US), and SMD and the Jarzynski equality (denoted as SMD/Jarzynski).

We studied dissociation of tri-N-acetyl-D-glucosamine (triNAG) from hen egg white lysozyme (LYZ) as our target. LYZ has long been studied as the ideal protein of many studies due to its antibacterial property[48]. triNAG binds to a cleft between two domains: a domain consisting of α helices (α domain) and a β-rich domain (β domain). Both experimental and computational studies indicated that the cleft can afford six N-acetyl-D-glucosamine (NAG) binding pockets from A to F, among of that, A-B-C is the main binding motif [49–52]. Recently, Zhong & Pastel used a polarizable force field, together with molecular mechanics with generalized Born and surface area (MM-GBSA), to investigate the A-B-C and B-C-D binding modes of triNAG to LYZ. However, neither of their models reproduced the binding free energy of the wild-type LYZ-triNAG complex[53]. The different of binding free energy is assumed coming from the neglecting of the contribution of the solvation free energy.

In this study, we show that the main interactions between LYZ and triNAG in the bound state agree with those found in the crystal structure. In addition, our estimation of binding free energy of the LYZ-triNAG complex is in agreement with experimental and the other computational results. Moreover, the combination of PaCS-MD and MSM allows the more cost-effective and accurate evaluation of binding free energy.

2. Calculation

We used the wide-type LYZ and triNAG (PDB ID: 1HEW) structure to genereate the simulation box. The initial box was 7.9094×7.66642×16.26561 nm^3 along the x, y, and z-axes, respectively, to accommodate for the large dissociation movement of triNAG along the z-axis (Fig 3a). Initially, the inter-COM distance between LYZ and triNAG was directed parallel to the z-axis. To avoid significant overall translation and reorientation of LYZ during triNAG dissociation, weak positional restraints were applied to the sulfur atoms of the cysteine residues involved in the four disulfide bonds of LYZ during the final stage of equilibration (step 5; see next paragraph) and in the production runs. As shown in Fig 3a, this system size was chosen so that the distances between the outermost atoms of the complex and the box edges were at least 1.5, 1.5, and 5.5 *nm* along the x, y, and z-axes, respectively. The box was solvated with TIP3P water and NaCl to ensure ionic concentration of $0.15\ M$ and charge neutrality. We used the AMBER99SB force-field[18] for LYZ and the GLYCAM06 force-field[54] for triNAG. All simulations were performed by GROMACS 5.0.5[55].

Simulation procedure with timestep 1 fs is carried out as followed. 1) Systems is performed steepest descent energy minimization followed by conjugate gradient method with heavy atom positional restraints with force constant of 1000 $kJ/mol\ nm^2$. 2) NVT ensemble annealing simulation is used to heat the system up from 0 K to 300 K within 500 ps, and thermostabilized at 300K for the next 500 ps. 3) Thermostabilizing simulation is switched to NPT ensemble for keeping pressure at 1.0 atm and temperature at 300K within 100 ps. Note that relaxation time of 0.1 ps for heat bath coupling and that of 2.0 ps for isotropic pressure coupling. 4) NPT ensemble equilibrium simulation continues for next 1 ns with the deduction of position restrained force constant 100 $kJ/mol\ nm^2$ every 100 ps until vanished. 5) For next 3.0 ns, simulation is carried out with positional restraints on the sulfur atoms of the cysteine residues (shown in yellow of Fig 3a). We used LINCS method to constrain the bond lengths[56] and leap-frog integration method[57] in steps 2-4, while velocity Verlet method[58] without bond constraints was taken advantage in step 5. Thermostat was performed by velocity rescaling[59] in steps 2-3 and a Nosé-Hoover method[60,61] in steps 4-5, while the used

barostat was Berendsen barostat[62] in step 3, a Parinello-Rahman barostat[63] in step 4, and a MTTK barostat[64] in step 5.

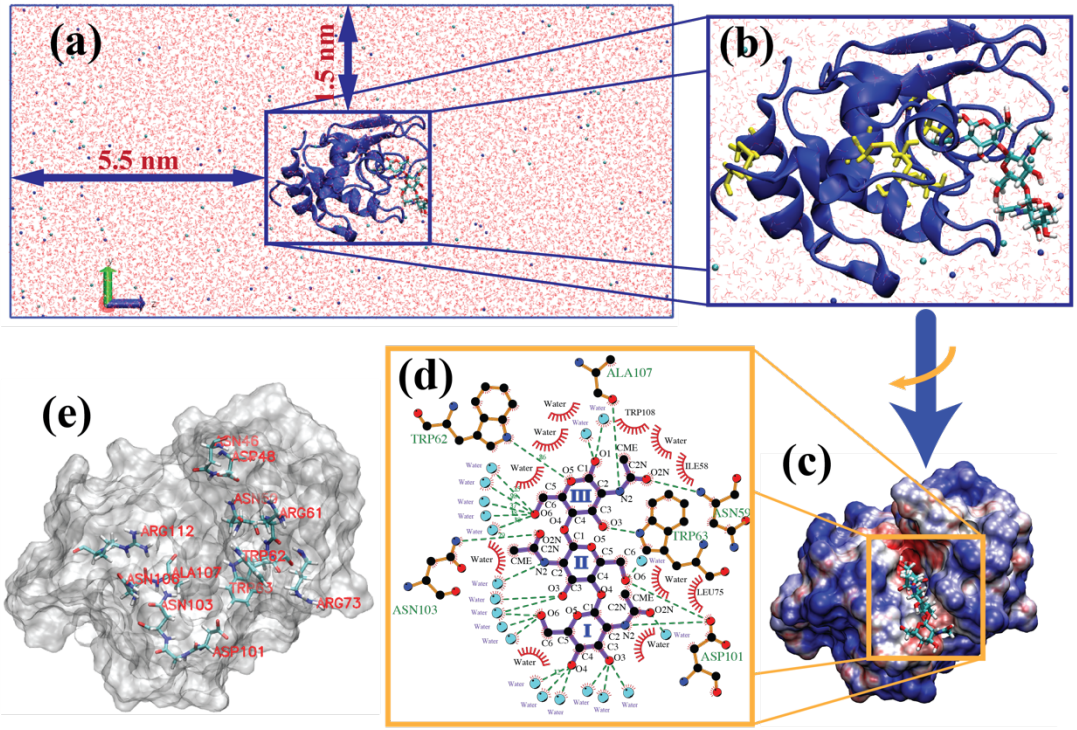


Fig 3. Visualization of the simulation box in the initial state and the key amino acid residues of LYZ that interact with triNAG after equilibration. (a) Overall arrangement of the LYZ-triNAG complex and solvent in the simulation box and (b) a close-up view. The residues shown by yellow Licorice models are disulfide-bonded cysteine residues and the molecule represented as a multicolored Licorice model is triNAG. (c) A view along the z-axis to show the electrostatic potential on the LYZ surface (blue: positive charges, red: negative charges). triNAG is shown as a Licorice model. (d) LigPlot+ diagram to show the interactions between LYZ and triNAG. Hydrophobic contacts are represented as spline curves outlining residue labels and hydrogen bonds are shown as dotted lines together with hydrogen bond distances. (e) Positions of the LYZ residues involved in hydrogen bonds with triNAG in the binding pockets. Blue and red residue labels show the residues situated in the α and β domains, respectively. Panels (a-c,e) and (d) were created by VMD[65] and LipPlot+[66], respectively.

3. Result and discussion

3.1. Interactions between LYZ and triNAG in the Bound State and the Stability of the Complex

Here we examine the interaction between LYZ and triNAG (Fig 3c and 3d)after performing relaxation simulation (step 5). The binding cleft of LYZ contains positively charged residues (ARG73 and ARG112), negatively charged residues (ASP52 and

ASP101), polar residues (GLN57, ASN59, and ASN103), and hydrophobic residues (LEU56, ILE58, TRP62, TRP63, LEU75, ILE98, ALA107, TRP108, and VAL109). It is interesting to note that negative charges (red in Fig 3c) are buried inside the binding cleft with the coverage of positive charges on the surface of LYZ (blue in Fig 3c) and hydrophobic residues. We found that within 1 μs conventional MD run, triNAG cannot dissociate from the complex of LYZ (as the RMSD and inter-COM distance (denoted as *d*) do not significantly change in Fig 4). In additions, Fig 4 shows inter-COM distance is stable while keeping long-lasting hydrogen bonds with ASN59, TRP62, TRP63, ASP101, ASN103, and ALA107 of LYZ, which is identical to the interaction in crystal structure PDB ID 1HEW[50] and by the other computational studies[51–53]. Those residues play important role in catalysis, binding affinity, and stability[67–70].

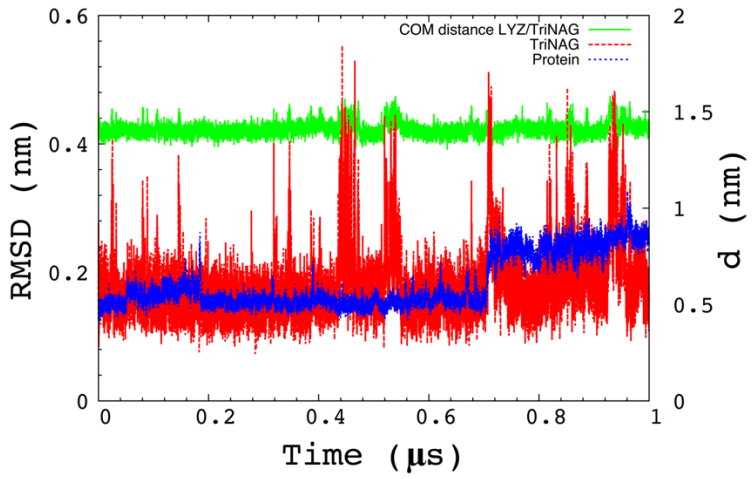


Fig 4. RMSD and COM distance in 1μs conventional MD simulation of the LYZ and triNAG complex.

Table 1. List of simulations and their conditions.

Simulation	n_{rep} / t_{cyc} (ns) in PaCS-MD or v in SMD (nm/ns)	# trials	MD time to reach 4 nm (ns)	MSM	US	Jarzynski
PaCS-MD$^{10,0.1}$	10 / 0.1	10	3.5 ± 1.0		Y (5.42)	
PaCS-MD$^{100,0.1}$	100 / 0.1	10	1.5 ± 0.2	Y (2.00)		
PaCS-MD^{10,1}	10 / 1.0	10	24.6 ± 10.1	Y (2.67)	Y (7.67)	
SMD^{fast}	1.25	24	2.0 ± 0.1		Y (12.12)	Y (0.12)
SMD^{med}	0.25	12	9.8 ± 0.4		Y (6.28)	Y (0.28)
SMD^{slow}	0.05	12	45.9 ± 9.3		Y(6.96)	Y (0.96)

3.2. LYZ-triNAG Dissociation by PaCS-MD

Although the LYZ-triNAG complex was stable during 1 μs, PaCS-MD can dissociate the complex very easily. We shows the time evolution of the largest inter-COM distance between LYZ and triNAG via the PaCS-MD simulation in Fig 5. For completing the dissociation of triNAG from LYZ ($d > 4\ nm$), it costs on average over trials at 34.8 $\pm$10.3 (3.48 ns), 14.5$\pm$1.7 (1.45 ns), and 24.6$\pm$10.1 cycles (24.6 ns), and the simulations were stopped at 41.6$\pm$10.4, 20.2$\pm$2.5, and 27.6$\pm$10.5 cycles for PaCS-MD$^{10,0.1}$, PaCS-MD$^{100,0.1}$, and PaCS-MD^{10,1}, respectively, when *d* reached *7 nm* (Fig 5 and Table 1). Compared to PaCS-MD$^{10,0.1}$, the number of cycles required for complete dissociation was reduced to 48.6 and 66.3% in PaCS-MD$^{100,0.1}$ and PaCS-MD^{10,1}, respectively. It is worthwhile to mention that although PaCS-MD$^{100,0.1}$ and PaCS-MD^{10,1} required the same computational resource per cycle, the sampling efficiency of PaCS-MD$^{100,0.1}$ was higher than that of PaCS-MD^{10,1} because of the former's fewer cycles to achieve complete dissociation. Moreover, the standard deviation of number of cycle of PaCS-MD$^{100,0.1}$ is also smaller resulting in the smaller variation between simulation lengths required for trials, as noted in Fig 5. In addition, the mechanism of PaCS-MD allows the increment of the probability observing rare event via number of replicas n_{rep} due to restarting MD simulations. Therefore, we claim that the increment of n_{rep} is better for sampling than that of simulation length t_{cyc}.

The dissociation process can be classified into three states: bound state, partially-bound state and unbound state. The bound state is defined as one in which the inter-COM distance increases slowly and almost linearly (the regions below the shaded regions in Fig 5). Next, the partially-bound state is defined as a state in which non-linear rapid increase of d occurs in which triNAG has few contacts left with LYZ (the shaded regions in Fig 5: $1.79 - 3.65$, $1.73 - 3.39$, and $1.82 - 3.18\ nm$ for PaCS-MD$^{10,0.1}$, PaCS-MD$^{100,0.1}$, and PaCS-MD^{10,1}, respectively). The unbound state is the regions above the shaded regions where *d* increases almost linearly and rapidly. We found that the average total number of cycles required for complete dissociation is mostly spent on the number of bound-state cycles, which were 24.1$\pm$11.2 (2.41 ns), 7.8$\pm$1.8 (0.78 ns), and 17.7$\pm$9.5 (17.7 ns) for PaCS-MD$^{10,0.1}$, PaCS-MD$^{100,0.1}$, and PaCS-MD^{10,1},

respectively. We defined the cycle with no significant increase of d as 'trapped' cycles. It occurred 8.8 ± 8.4 and 6.6 ± 6.34 times on average in PaCS-MD$^{10,0.1}$ and PaCS-MD^{10,1}, respectively, whereas there are no trapped cycles in PaCS-MD$^{100,0.1}$. Traps mostly occurred in the bound and partially-bound states. The average number of continuous trapped cycles was 4.4 ± 2.3 and 3.8 ± 2.9 for PaCS-MD$^{10,0.1}$ and PaCS-MD^{10,1}, respectively. This again indicates that PaCS-MD larger n_{rep} increase the efficience of sampling.

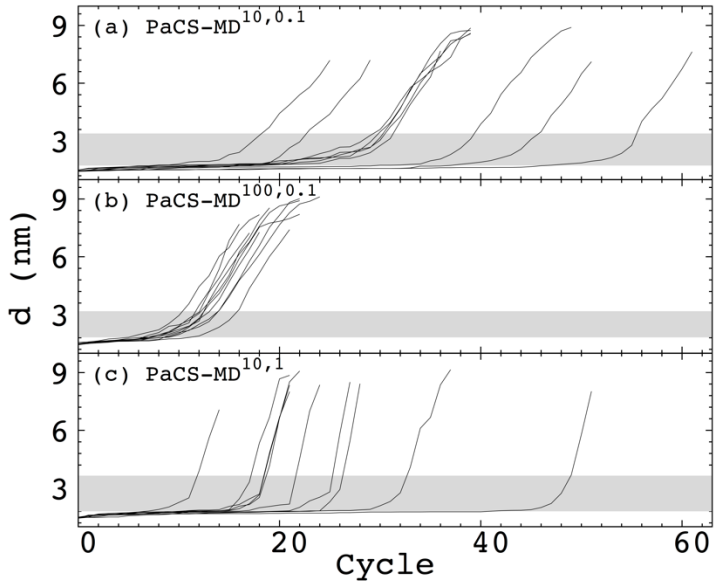


Fig 5. Evolution of the inter-COM distance between lysozyme and triNAG, d, in the top reactive trajectories during each PaCS-MD trial as a function of the number of cycles for (a) PaCS-MD$^{10,0.1}$, (b) PaCS-MD$^{100,0.1}$, and (c) PaCS-MD^{10,1}. The meanings of the shaded regions are marked for the partially-bound state.

For comparison with SMD pulling speed, we also estimate the movement speed of triNAG according to the dissociation process. In the unbound state, the average speeds of triNAG movement were 0.64 ± 0.08, 0.77 ± 0.16 and 1.56 ± 0.27 $nm/cycle$ for PaCS-MD$^{10,0.1}$, PaCS-MD$^{100,0.1}$, and PaCS-MD^{10,1}, which correspond to 6.4, 7.7 and 1.6 nm/ns, respectively. This speed in the PaCS-MD^{10,1} simulation was equivalent to the pulling velocity of SMD^{fast} (1.25 nm/ns), while those of PaCS-MD$^{10,0.1}$ and PaCS-MD$^{100,0.1}$ were 5 times faster than the pulling velocity of SMD^{fast}.

3.3. Effects of Velocity Re-randomization and Selection on triNAG Dissociation during PaCS-MD

We examined diffusive properties during PaCS-MD by inspecting the self-diffusion constant D by the Einstein relation:

$$D = lim_{t\to\infty} \frac{\langle \Delta r^2(t) \rangle}{6t} \quad (15)$$

Equation (15) implies that D can be calculated by performing least squares fitting of the time evolution of mean square displacement (MSD) $\langle \Delta r^2(t) \rangle$ to a straight line. We first calculated the effective diffusion in "reactive trajectories", which were used for the initial structures for US. Due to different behavior of each states, they was analyzed separatedly. The obtained $\langle \Delta r^2(t) \rangle$ is shown in Fig 6. The self-diffusion constants in the unbound state, $D_{unbound}^{react.}$, were $(6.6 \pm 1.9)\times 10^{-5} cm^2/s$, $(7.7 \pm 1.2)\times 10^{-5} cm^2/s$, and $(1.9 \pm 0.8)\times 10^{-5} cm^2/s$ for PaCS-MD$^{10,0.1}$, PaCS-MD$^{100,0.1}$, and PaCS-MD^{10,1}, respectively, which indicates that shorter t_{cyc} (= 0.1 ns) accelerated effective diffusion more than threefold compared to t_{cyc} = 1 ns. The values of $D_{unbound}^{react.}$ obtained from PaCS-MD simulations were significantly larger than triNAG's free diffusion constant, $1.1 \pm 0.5\times 10^{-5} cm^2/s$, confirming that PaCS-MD enhanced the effective diffusion constants of the unbound state. In addition, the reactive trajectories are continuous in conformational space but might be discontinuous in phase space. Hence, we can conclude that the velocity re-randomization causes perturbation of the reactive trajectories at each concatenating point in the trajectory.

The length of the fragment of trajectories Δt_{frag} contributing to the reactive trajectories is also of interest. If Δt_{frag} is too short, the system might not relax sufficiently after velocity re-randomization. However, the Δt_{frag} values for PaCS-MD$^{10,0.1}$ and PaCS-MD^{10,1} were 79.2 ± 7.7 and $842.4 \pm 122.8\ ps$, respectively, which are significantly longer than the safe limit exchange time interval 4 ps in temperature REMD[71].

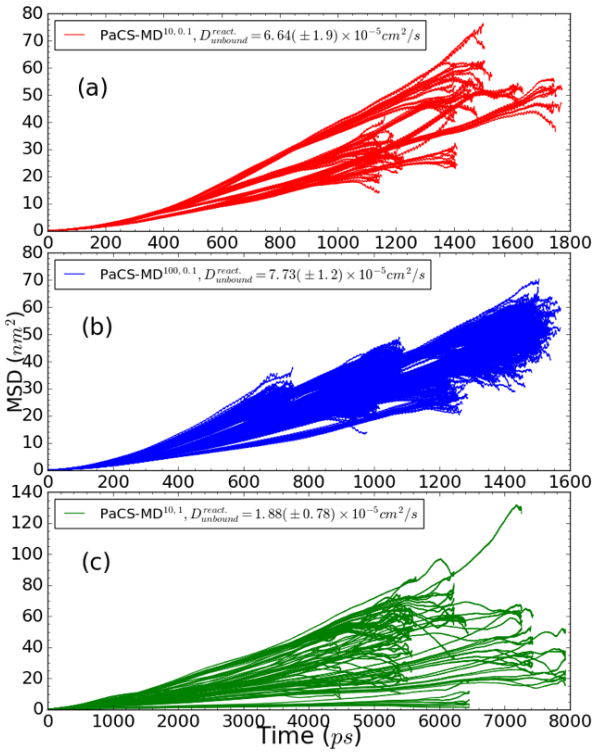


Fig 6. Mean-square displacement (MSD) of triNAG calculated from the reactive trajectories for a) PaCS-MD$^{10,0.1}$, b) PaCS-MD$^{100,0.1}$ and c) PaCS-MD^{10,1}.

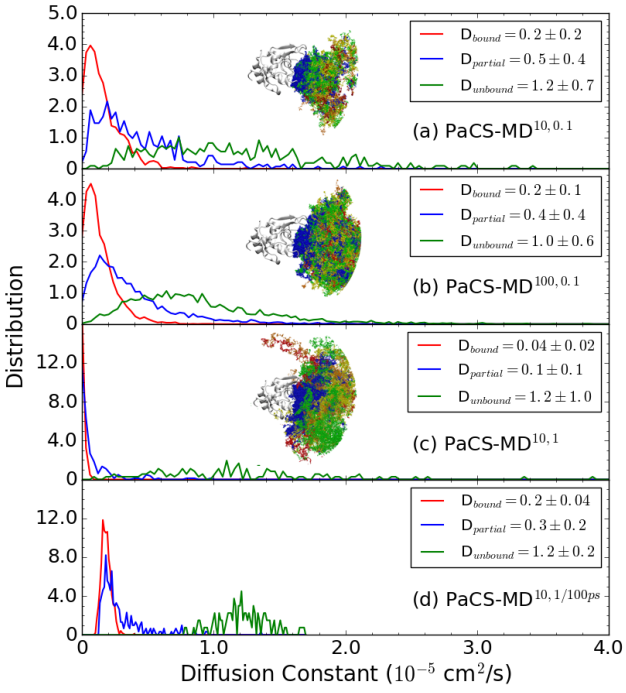


Fig 7. Distributions of the triNAG self-diffusion constants, D_{bound} (red), $D_{partial}$ (blue), and $D_{unbound}$ (green) in (a) PaCS-MD$^{10,0.1}$, (b) PaCS-MD$^{100,0.1}$, (c, d) PaCS-MD^{10,1} shown as probability densities. The densities in (c) and (d) were calculated from entire 1 ns and the first 0.1 ns trajectories, respectively. Insets show each COM trajectory of triNAG around LYZ (white cartoon model) in different colors, depending on the values of the diffusion constant: blue ($D \leq 0.5$; all units 10^{-5} cm^2/s), green ($0.5 < D \leq 1.0$), yellow ($1.0 < D \leq 1.5$), orange ($1.5 < D \leq 2.0$) and red ($2.0 < D$). The values after ± show standard deviations.

To evaluate the influence of velocity rerandomization in trajectories, we analyzed self-diffusion constants of tri-NAG in bound, partially-bound and unbound states (D_{bound}, $D_{partial}$, and $D_{unbound}$), as depicted in Fig 7. We add the calculation for the first 0.1 ns of PaCS-MD^{10,1} in Fig 7d for comparison with PaCS-MD$^{10,0.1}$. We found the same trend of the distribution of diffusion constants as of random walk trajectories generated by varying the concentration of random point obstacles[72]. In addition, $D_{unbound}$ (1.2 ± 0.2, 1.0 ± 0.6 and $1.2 \pm 0.7\times10^{-5} cm^2/s$ for PaCS-MD$^{10,0.1}$, PaCS-MD$^{100,0.1}$, and PaCS-MD^{10,1}) are in good agreement with the free diffusion constant ($1.1 \pm 0.5\times10^{-5} cm^2/s$) that we calculated. If velocity re-randomization has a significant effect on diffusion, the diffusion coefficient will depend on the simulation length (1.0 or 0.1 ns), this would lead to the effect on diffusion of velocity rerandomization. However, we found the effect of velocity re-randomization is weak in the unbound state, and no significant influence on the diffusion constants was observed . However, D_{bound} is the same for the first 0.1 ns (Fig 7(a,b,d)) but is smaller for 1.0 ns (Fig 7(c)), indicating the effect of velocity re-randomization on diffusion depends on the length of the trajectory in the bound state. Moreover, D_{bound} and $D_{partial}$ (red and blue curves in Fig 7) are mainly populated below $0.5\times10^{-5} cm^2/s$ and spatially form a low mobility region around the binding pockets (the trajectories shown by blue in the insets of Fig 7). As triNAG dissociates farther, the higher imhomogeneous mobility regions were observed but were not through the clear color variations in Fig 7. Broader range of $D_{unbound}$ than that of the other states shows the imhomogeneous of mobility in the unbound state. Specifically, $D_{unbound}$ of PaCS-MD$^{10,0.1}$, PaCS-MD$^{100,0.1}$, and PaCS-MD^{10,1} is significantly larger than D_{bound}, by 5.8, 5.3, and 20.5 times, respectively.

Interestingly, the smaller value of D_{bound} in PaCS-MD^{10,1} obtained from full 1.0 ns trajectories (Fig 7(c)) compared to that obtained from the first 0.1 ns indicates that a longer MD simulation time did not accelerate diffusion in the bound state. These results suggest that velocity re-randomization enhanced sampling in the bound state. To shed light on the effect of velocity re-randomization on selection, we analyzed the time evolution of the probability of selection of the selected snapshots (Fig 8). Bound state selected snapshots tends to be near the beginning of each MD, indicating that velocity

re-randomization enhances movements leading toward dissociation in the PaCS-MD scheme with quick decaying. Similar tendencies were observed in the partially-bound states during PaCS-MD$^{10,0.1}$ and PaCS-MD^{10,1} but not during PaCS-MD$^{100,0.1}$. The exceptional case for PaCS-MD$^{100,0.1}$ is from no trapped cycle observed in the bound and partially-bound states of the structures than that of PaCS-MD$^{10,0.1}$ and PaCS-MD^{10,1}. If a significant increase in *d* was not observed, snapshots near the beginning of the MD run were selected, which raised the probability of snapshot selection from this time region. The selected unbound state snapshots located near the end of the MD run frequently because the movement of triNAG in the unbound state is largely determined by diffusion, and larger deviations should occur near the end of the MD simulation in a diffusion-dominant environment.

For summary, results reported above imply that dynamics behavior in individual short simulations within PaCS-MD scheme is the same as expected in unbiased MD simulations. Thus, we judged that MSM can be appropriately applied to PaCS-MD trajectories.

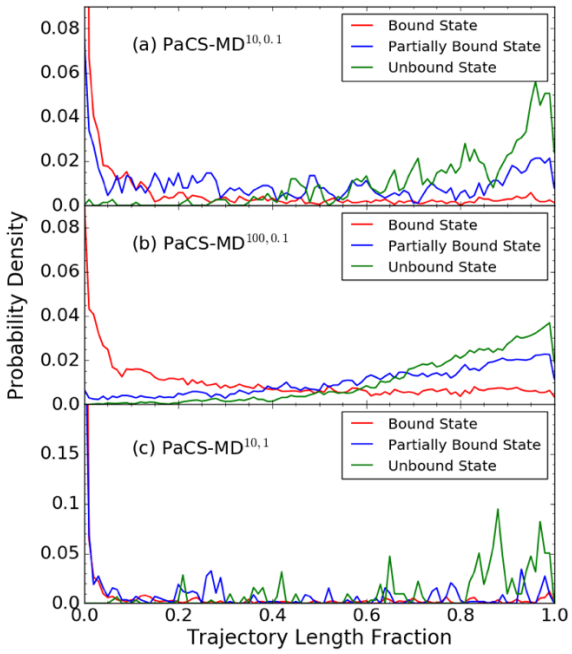


Fig 8. Probability of selection as a fraction of time of the selected snapshots versus the total length of each MD (1.0 or 0.1 ns) in each state for (a) PaCS-MD$^{10,0.1}$, (b) PaCS-MD$^{100,0.1}$, and (c) PaCS-MD^{10,1} in the bound (red), partially-bound (blue), and unbound (green) states.

3.4. LYZ-triNAG Dissociation by SMD

In SMD, ligand dissociation was induced by the steering force. Figure 5 shows the time evolution of the inter-COM distance, d, and the force between LYZ and triNAG as a function of SMD time. Unlike PaCS-MD, the time evolution of d exhibited a steep jump between two linear regions (Fig 9(a-c)). The initial linear regions range from 1.3 to 1.7 nm in all three cases. The jump started when the steering force became maximum. The average maximum forces during the dissociation process were 336.1 ± 44.5, 303.9 ± 43.6, and 267.2 ± 33.5 $kJ/mol \cdot nm$ in the SMD^{fast}, SMD^{med}, and SMD^{slow} simulations, respectively, which are in the same range as the AFM disruption forces previously reported[73]. The lower the pulling rate, the weaker the maximum force required to dissociate triNAG. After the steep jump, d linearly increased and the force converged toward zero at around 2.2–2.6 nm. We found two different patterns in the steering force as a function of d: single peak, and double peaks shown in Table 2. We show the values in parenthesis in Table 2 the number of cases in which the heights of the two peaks are the same (so-called same-height double peaks). We found that the heights of the first and second peaks decreased as the SMD velocity decreased. The force peak for the single-peak cases was generally larger than that of the double-peak cases while the heights of the same-height double peaks were lower than the single peak by 100 kJ/mol.nm. After reaching the first peak, triNAG quickly dissociated from LYZ; however, triNAG remained trapped in the double-peak cases, which correspond to small plateau regions (red lines in Fig 9). The standard deviation of the positions of the first peaks were always small (≤ 0.1 nm), showing that the first stage of the dissociation processes in SMD started from the same position.

Table 2. Characteristic inter-COM distances and forces in SMD.

Sim.	Type	#	1st peak		2nd peak		Convergence
			d (nm)	Force (kJ/mol nm)	d (nm)	Force (kJ/mol nm)	d (nm)
SMD^{fast}	Single	13	1.5 ± 0.1	351.8 ± 38.7	-	-	2.6 ± 0.2
	Double	11(1)	1.5 ± 0.0	321.9 ± 40.0	2.2 ± 0.3	182.1 ± 35.1	2.6 ± 0.2
SMD^{med}	Single	5	1.5 ± 0.1	309.4 ± 46.6	-	-	2.5 ± 0.1
	Double	7(1)	1.5 ± 0.1	292.7 ± 41.8	2.4 ± 0.5	159.4 ± 24.9	2.4 ± 0.1
SMD^{slow}	Single	8	1.5 ± 0.1	269.4 ± 37.2	-	-	2.2 ± 0.2
	Double	4(0)	1.4 ± 0.0	247.9 ± 11.5	2.0 ± 0.3	149.7 ± 5.9	2.3 ± 0.3

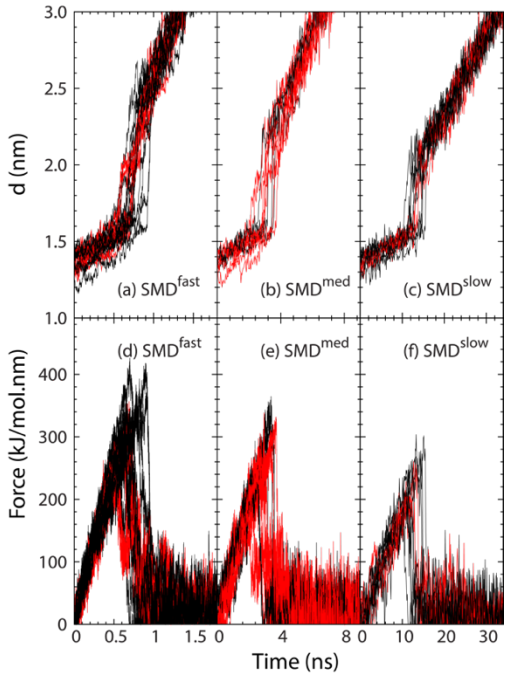


Fig 9. (a-c) Time evolution of the inter-COM distance, d, and (d-f) force between LYZ and triNAG as a function of SMD time for (a,d) SMD^{fast}, (b,e) SMD^{med}, and (c,f) SMD^{slow}. The red lines show the cases in which small plateau regions are seen in (a-c). In these cases, double force peaks as a function of time were observed (also see Table 3)

3.5. Dissociation Pathways in PaCS-MD and SMD

We analyzed the spatial distribution of the dissociation pathways to obtain better insight into the relation between the dissociation pathway and free energy. Fig 10(a) shows the COM positions of triNAG along 10 representative reactive trajectories, each of which is the top ranked reactive trajectory in each PaCS-MD[10,0.1] trial. The inset of Fig 10(a) depicts the triNAG COM positions in all PaCS-MD trajectories generated in one trial. Interestingly, a set of trajectories in PaCS-MD generated in one trial formed a barbed zigzag rod connecting the bound and completely unbound states, as shown in the inset of Fig 10(a). We introduced an effective diameter $\sigma^{each}(d)$ of a cross section of trajectories as a function of the inter-COM distance d quantity for better insight in sampling efficiency. $\sigma^{each}(d)$ is defined as:

$$\sigma^{each}(d) = \sqrt{\frac{4S^{each}(d)}{\pi}} \tag{16}$$

where $S^{each}(d)$ is a cross section of the sampled triNAG COM positions at d. The average of $\sigma^{each}(d)$ over trials, $\overline{\sigma^{each}}(d)$, is shown in Fig 10(b). $\overline{\sigma^{each}}(d)$ is larger in

the bound state (d < 2 nm) but is almost flat after complete dissociation (d > 2.5 nm). This reflects that more cycles were spent in the bound state than in the unbound state, resulting in larger $\overline{\sigma^{each}}(d)$ values in the bound state. The plateau value of $\bar{\sigma}$ in the unbound state is consistent with triNAG diffusing essentially freely in the unbound state. In the unbound state, the average values of $\overline{\sigma^{each}}(d)$ were $0.96 \pm 0.33, 1.50 \pm 0.74$ and $2.07 \pm 1.31\ nm$ for PaCS-MD$^{10,0.1}$, PaCS-MD$^{100,0.1}$, and PaCS-MD^{10,1}, respectively. This shows that a longer simulation time for each replica provides a larger sampling diameter compared to increasing the number of replicas. In the partially-bound state, the sampling diameter is comparable between PaCS-MD$^{100,0.1}$ and PaCS-MD^{10,1}.

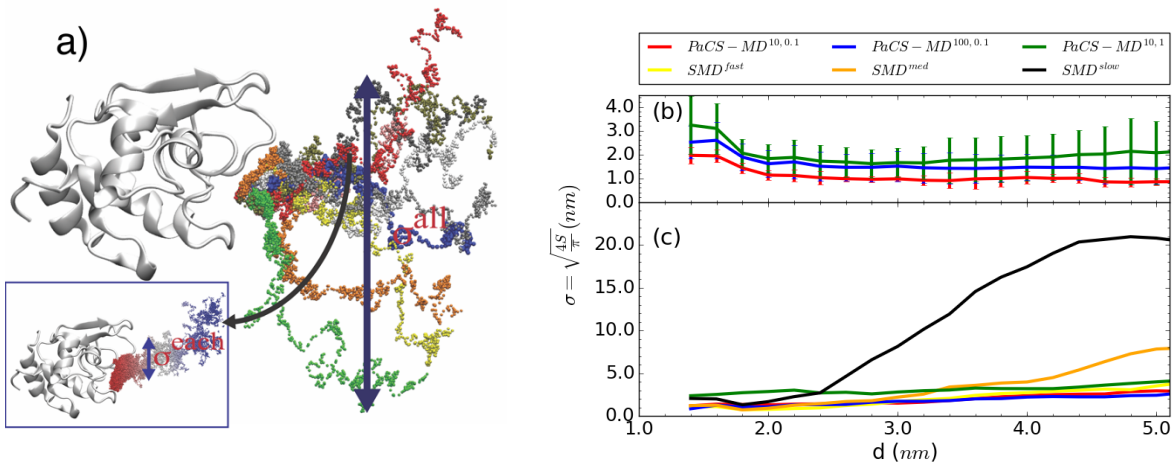


Fig 10. (a) Dissociation pathways of triNAG represented by the COM positions of triNAG (small spheres) in the first reactive trajectories of 10 PaCS-MD$^{10,0.1}$ trials from LYZ (white cartoon model). Inset shows all the trajectories generated in a representative representative trial of PaCS-MD$^{10,0.1}$ (red in main panel). (b) Effective diameter σ of the sampled area per trial of PaCS-MD as a function of the inter-COM distance d*. (c) Effective diameter σ over all trials. The inset shows a close up. PaCS-MD$^{10,0.1}$ (red), PaCS-MD$^{100,0.1}$ (blue), PaCS-MD^{10,1} (green), SMD^{fast} (magenta), SMD^{med} (orange), and SMD^{slow} (black). Error bars show standard deviations.*

We also examined the variation of the dissociation pathways generated by distinct PaCS-MD and SMD trials (Fig 12). This figure clearly shows that significantly different dissociation pathways are generated in each type of simulation. To quantify this variation, we also calculated the $\sigma(d)$ for the PaCS-MD reactive trajectories of all trials and all SMD trajectories (Fig 10(a)), denoted as $\sigma^{all}(d)$, and the results are shown in Fig 10(c). In the bound state, the average values of $\sigma^{all}(d)$ were 1.03 ± 0.08 and

$1.06 \pm 0.15\ (nm)$ in the PaCS-MD$^{10,0.1}$ and PaCS-MD$^{100,0.1}$ simulations, respectively, and are comparable to the values 0.91 ± 0.17 and $1.03 \pm 0.41\ (nm)$ obtained by SMD^{fast} and SMD^{med}, respectively. However, $\sigma^{all}(d)$ obtained by SMD^{slow} and PaCS-MD^{10,1} for the bound state were significantly larger, 1.76 ± 0.29 and $2.61 \pm 0.17\ (nm)$, respectively. In the unbound state, $\sigma^{all}(d)$ obtained by SMD^{med} and SMD^{slow} steeply increased as d increased. We note that diffusion governs the movement along the x and y directions because a pulling force was applied only along the z direction. Therefore, the ratio of SMD^{slow} simulation time spent at $d = 4\ nm$ versus SMD^{med} is 4.7, consistent with the ratio of $\sigma^{all}(d = 4)$, 4.4 (Fig 10(c) and Table 1).

3.6. Dissociation Free Energy

The free energy profile (potential of mean force, PMF) of triNAG dissociation from LYS as a function of the inter-COM distance was calculated by combinations of PaCS-MD and MSM (PaCS-MD/MSM), PaCS-MD and US (PaCS-MD/US), SMD and US (SMD/MS), and SMD and the Jarzynski equation (SMD/Jarzynski) (shown in Fig 11). Since the free energy profiles were obtained as the average over distinct dissociation pathways (shown in Fig 12), they should be clearly distinguished from the minimum free energy path. The free energy profiles were all flat in the inter-COM distance range 4.0–4.5 nm, that help us to define the dissociation free energy ΔG_d as the energy difference between the bound state. We assumed that the calculated dissociation free energies are equal to the negative value of the binding free energy ΔG_b as $\Delta G_b = -\Delta G_d$.

In PaCS-MD/MSM, a MSM was constructed using PaCS-MD trajectories generated by PaCS-MD$^{100,0.1}$ and PaCS-MD^{10,1} simulations (Table 1). Note that the MSM was built using all trajectories of each trial and the average PMF was obtained over all trials, as shown in Fig 11(a). Each trial of PaCS-MD$^{10,0.1}$ lacked adequate statistics to build the MSM properly. After careful evolution of the number of microstates and the implied time scale as a function of lag time τ, we determined 50 microstates for both cases and selected 45 and 305 ps as the best τ values for PaCS-MD$^{100,0.1}$ and PaCS-MD^{10,1}, respectively. These values were much shorter than values typically used in MSMs constructed from microsecond trajectories for

folding/unfolding studies. However, the fraction of the total simulation time in each trial versus τ is approximately 10^4, consistent with the value suggested for enhanced sampling methods to achieve convergence of MSMs[36]. Moreover, Suarez *et al.* conducted detailed analysis of MFPT with non-Markovian estimators and found a reduction in the bias intrinsic to Markov MFPT estimation, even at the shortest lag times or simple discretization of the configuration in one dimensional space[74]. In addition, Zhang *et al.* used 0.5 ps for τ, which is much shorter than our value, to build transition matrices for MSMs from replica exchange simulation[75]. Therefore, we believe our choice of τ is reasonable. The obtained dissociation free energies were 27.8 ± 0.8 and $30.5 \pm 0.8\ kJ/mol$ for PaCS-MD^{10,1} and PaCS-MD$^{100,0.1}$, respectively (Table 3), which are comparable to the binding free energy values measured by isothermal titration calorimetry (ITC) and surface plasmon resonance (SPR)[51,76].

To build a MSM directly from PaCS-MD trajectories, we should carefully choose n_{rep} and $\mathrm{t_{cyc}}$. Indeed, to build a proper MSM also requires the correct choice of $\mathrm{t_{cyc}}$ i.e. $\mathrm{t_{cyc}}$ affects the choice of lag time τ and the definition of microstates in MSM because the condition $\tau \leq \mathrm{t_{cyc}}/2$ is expected from the statistical point of view[44]. Therefore, the merit of using longer τ is to provide flexibility in deciding proper τ and microstates. In this study, as mentioned above, $\mathrm{t_{cyc}}$ = 0.1 ns in PaCS-MD$^{100,0.1}$ simulations was sufficient to build a MSM. It should be noted that we obtained similar ΔG_d values from PaCS-MD$^{100,0.1}$ and PaCS-MD$^{10,0.1}$ simulations despite using different lag times. Using PaCS-MD to generate initial pathways with a limited number of n_{rep}, followed by MD simulations to obtain more statistics to build a MSM, provides more options. In this case, PaCS-MD can be mainly dedicated to accelerating expected movements with sufficiently short MDs. A MSM can be built by extending the MD simulation length and/or by adding more MD simulations later. Also, as in the case where we noticed insufficient statistics to build a MSM after PaCS-MD, more MD trajectories can be added to properly construct a MSM.

We also calculated the free energy profile using PaCS-MD/US (Fig 11(b)). Sharp peaks observed at around 1.6 nm were not observed in the PaCS-MD/MSM results. While the free energy profile in PaCS-MD/MSM was directly calculated from PaCS-

MD trajectories, that in PaCS-MD/US was obtained by additional multiple umbrella samplings with restrained inter-COM distances. The peak at $1.6\,nm$ indicates the region where triNAG was trapped in the binding pockets and strongly corresponds to the first peak of steering forces in SMD. The minimum at $1.9\,nm$ after the first peak is related to the upper limit of the bound state, as shown in Fig 5. PaCS-MD^{10,1}/US yielded a free energy difference larger than the experimental values, while PaCS-MD$^{10,0.1}$/US gave a more reasonable value, $26.8 \pm 1.3\,kJ/mol$.

The energy profiles of SMD/US indicated similar tendencies to those of PaCS-MD/US, including the position of the minimum at $1.9\,nm$ (Fig 11(c)) This minimum was clearly seen in SMD^{fast} and SMD^{med} simulations, but was not observed in SMD^{slow} simulations. The binding free energies obtained by SMD^{fast}/US ($29.7 \pm 0.9kJ/mol$) and SMD^{med}/US ($27.2 \pm 1.1kJ/mol$) are in the range of available experimental results (Table 3), whereas SMD^{slow}/US under-estimated ΔG_b.

SMD/Jarzynski (Fig 11(d)) significantly overestimated the dissociation free energy between LYZ and triNAG. Yamashita and Fujitani showed that SMD/US yielded higher free energy difference than using the combination of US with multi-step targeted MD[29]. However, we did not observe such overestimation in SMD/US in our case, probably because of the size of small ligand leading to the restoring during US. The dissociation free energy profiles obtained from SMD/Jarzynski showed a clear tendency that lower SMD pulling velocities resulted in lower binding free energy (Fig 11(d)). The first plateaus around 1.5–2.1 nm correspond to the region between the first and second force peaks shown in Table 2. The heights of the plateaus were 93.3 (SMD^{fast}), 46.9 (SMD^{med}), and 20.5 kJ/mol (SMD^{slow}). The dissociation free energies obtained by SMD/Jarzynski (Table 3) were larger by factors of 2–5 than that expected from experimental results[51]. The steering force in SMD induced a biased dissociation process, which might significantly change the dissociation free energy.

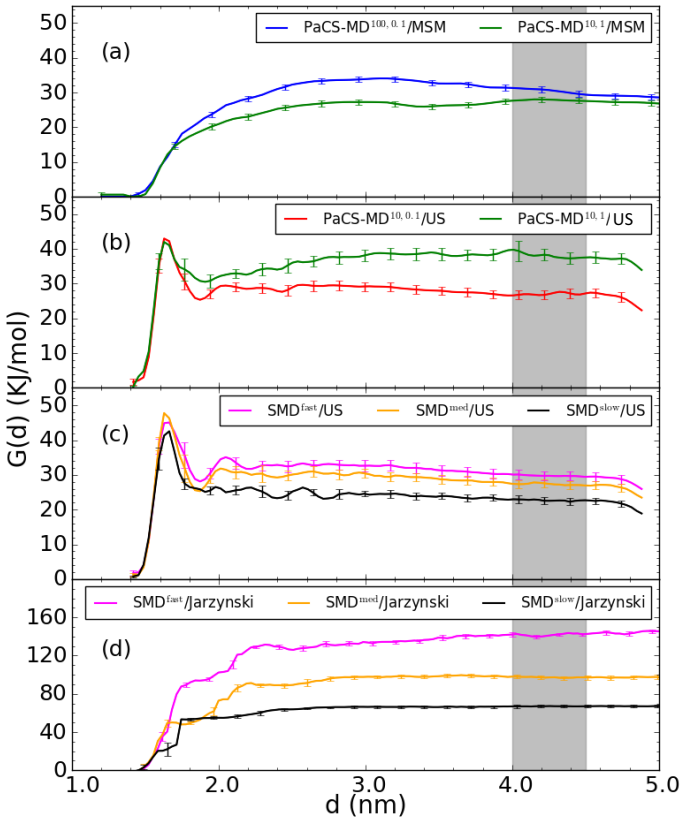


Fig 11. Dissociation free energy profiles calculated by combinations of (a) PaCS-MD and MSM, (b) PaCS-MD and US, (c) SMD and US and (d) SMD and the Jarzynski equality as functions of the inter-COM distance d. *Error bars show standard errors of the mean. The average values over the gray shaded regions (4.0 – 4.5 nm) were considered as the dissociation free energy* ΔG_d.

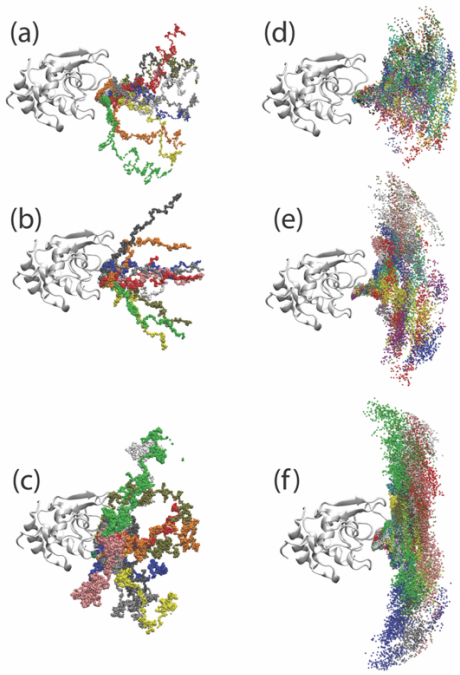


Fig 12. Snapshots of the COM of triNAG in all trajectories. Color differences is used for marking that the COMs of triNAG are in the same trial of (a) PaCS-MD10,0.1, (b) PaCS-MD100,0.1, (c) PaCS-MD10,1, (d) SMDfast, (e) SMD med, (f) SMDslow.

Table 3. Comparison of dissociation and association free energies obtained by simulations and experiments.

Free energy difference	Methods	Free energy[a] (kJ/mol)
ΔG_d	PaCS-MD^{10,1}/MSM	27.8 ± 0.8 (27.6 ± 0.8)
	PaCS-MD$^{100,0.1}$/MSM	30.5 ± 0.8 (31.3 ± 0.8)
	PaCS-MD^{10,1}/US	37.7 ± 1.9 (39.8 ± 3.2)
	PaCS-MD$^{10,0.1}$/US	26.8 ± 1.3 (26.6 ± 1.2)
	SMD^{fast}/US	29.7 ± 0.9 (30.1 ± 0.9)
	SMD^{med}/US	27.2 ± 1.1 (27.4 ± 1.1)
	SMD^{slow}/US	22.6 ± 0.9 (23.0 ± 1.0)
	SMD^{fast}/Jarzynski	148.1 ± 2.3 (142.1 ± 2.1)
	SMD^{med}/Jarzynski	97.5 ± 2.2 (98.0 ± 2.0)
	SMD^{slow}/Jarzynski	67.4 ± 1.5 (66.9 ± 1.5)
ΔG_b	ITC at pH 4.6[51]	-29.2
	ITC at pH 7.3[51]	-28.5
	SPR at pH 7.4[51]	-26.9
	ITC at pH 7.0[76]	-28.9

3.7. Disruption of LYZ-triNAG Interactions during the Dissociation Process

We examined the dissociation process of triNAG from LYZ by analyzing the breakage of the intermolecular hydrogen bonds shown in Fig 3(d,e) and determined the order in which LYS residues dissociated from triNAG in each trial (Table 4 and Table 5). There are 5.6 hydrogen bonds in average between LYZ and triNAG in the bound state during 1 μs conventional MD and additional transient hydrogen bonds were formed in the partially-bound state. The number of hydrogen bonds between LYZ and triNAG during dissociation in PaCS-MD$^{10,0.1}$, PaCS-MD$^{100,0.1}$, and PaCS-MD^{10,1} simulations were on average 8.1, 8.7, and 10.6, and those in SMD^{fast}, SMD^{med}, and SMD^{slow} simulations were 7.5, 9.4, and 12.3, respectively. These results indicate that the longer the MD run, or the slower the SMD velocity, the larger the number of hydrogen bonds formed during dissociation. We found that ASN59, TRP62, TRP63, and ALA107 are key residues that always formed hydrogen bonds with triNAG in the bound state. Interestingly, the TRP62 and TRP63 hydrogen bonds with NAG III located in pocket C broke before

those of ASN59 and ALA107 with NAG II situated in pocket C, regardless of the simulation type. ASN59, TRP62 and TRP63 belong to the β domain, while ALA107 is situated on the α domain and is essentially outside the cleft (Fig 3(d,e)). NAG II, which binds to ASN103 and ASP101 of pocket B, dissociated after NAG III. triNAG did not slide along the binding pockets during dissociation but rather dissociated perpendicularly from the binding cleft starting from NAG III. If such sliding indeed happens, it should simultaneously break all the inter-molecular hydrogen bonds, which should suddenly increase the free energy. The order of hydrogen bond breakage during SMD depended on the pulling velocity, however, no clear variations were seen during PaCS-MD. SMD can lead to dissociation orders different from those observed in PaCS-MD. ASP48 is situated farther and deeper in the binding cleft compared to ASN59, and rarely forms hydrogen bonds with triNAG in the equilibrium bound state. However, during dissociation, ASP48 frequently formed hydrogen bonds with triNAG and then broke during SMD^{med} (9/12 trials) and SMD^{slow} (11/12 trials). In $PaCS\text{-}MD^{10,1}$ simulations, ARG112 formed transient hydrogen bonds with triNAG, and ASN106 played the same role in $PaCS\text{-}MD^{100,0.1}$ simulation. ASN106 and ARG112 are located on the surface of the α domain. The hydrogen bond between ARG73 and NAG III tended to break first during PaCS-MD, whereas that between ARG61 and NAG I was the first hydrogen bond lost during SMD. These hydrogen bonds were both formed during dissociation. In PaCS-MD simulation, NAG I tended to dissociate first and was exposed to water before NAG III dissociation, whereas these events were reversed during SMD. In PaCS-MD simulation, triNAG dissociation started from NAG I, then NAG III, and finally NAG II. The difference in dissociation order in SMD simulation might be due to force bias, which may contribute to a higher dissociation free energy in SMD^{fast}/Jarzynski.

Table 4. Residual order of hydrogen bond breakage with triNAG in PaCS-MD.

Trajectory	Hydrogen bonds of triNAG broken order												
	1	2	3	4	5	6	7	8	9	10	11	12	13
$PaCS\text{-}MD^{10,0.1}1$	ARG61	TRP62	ARG73	TRP63	ASN59	ASP101	ALA107						
$PaCS\text{-}MD^{10,0.1}2$	TRP62	TRP63	ASN103	ASN59	ALA107	ASP101	ASN106						
$PaCS\text{-}MD^{10,0.1}3$	ARG73	TRP62	TRP63	ASP101	ASN59	ALA107	ASN103	ASN106					
$PaCS\text{-}MD^{10,0.1}4$	ASP101	ARG73	TRP63	TRP62	ASN59	ASN46	ALA107	ASN103	ASN106	ARG112	GLY102		
$PaCS\text{-}MD^{10,0.1}5$	ARG73	TRP62	TRP63	ASP48	ALA107	ASN59	ASP101	ASN103	GLY102				

PaCS-MD$^{10,0.1}$6	ARG73	TRP62	ASP101	TRP63	ASN103	ALA107	ASN59	ASN106					
PaCS-MD$^{10,0.1}$7	TRP62	TRP63	ARG61	ASN59	ASP101	ALA107	ASN103	GLY102					
PaCS-MD$^{10,0.1}$8	ASP101	ARG73	GLY102	TRP62	TRP63	ASN103	ALA107	ASN59					
PaCS-MD$^{10,0.1}$9	TRP62	TRP63	ALA107	ARG73	ASN59	ASN103	ASP101						
PaCS-MD$^{10,0.1}$10	ARG73	ARG61	TRP62	ASP101	TRP63	ASN59	ALA107	ASN103					
PaCS-MD^{10,1}1	ARG73	TRP62	TRP63	ASP101	ASN59	ALA107	ASN103	GLY102	ASN106	ARG112			
PaCS-MD^{10,1}2	ARG73	TRP62	ASP101	TRP63	ARG61	ASN59	ALA107	ASN103	ARG112	GLY102			
PaCS-MD^{10,1}3	ARG73	TRP62	TRP63	ASN59	ALA107	ASN46	ASP101	ARG112	ASN103	ASN106	GLY102		
PaCS-MD^{10,1}4	ARG73	ARG61	TRP62	TRP63	ASP101	ASN59	ALA107	ASN103	ARG112	GLY102	ASN106		
PaCS-MD^{10,1}5	ARG73	ASP101	TRP63	TRP62	ASN59	ALA107	ASN103	GLY102	ASN106	ARG112			
PaCS-MD^{10,1}6	TRP62	ASP48	ARG73	TRP63	ASN59	ALA107	ARG112	ASN106	ASN103	ASP101	GLY102		
PaCS-MD^{10,1}7	ASP101	TRP62	ARG73	TRP63	ASN59	ALA107	ASN103	ASN46	GLY102	ASN106	ARG112		
PaCS-MD^{10,1}8	ASP48	TRP62	ARG112	ALA107	ASN59	ASN46	TRP63	ASN103	ASP101	ARG73	ARG61	GLY102	
PaCS-MD^{10,1}9	TRP62	ASP48	ARG61	ARG73	TRP63	ASN59	ALA107	ASP101	GLY102				
PaCS-MD^{10,1}10	ARG73	ASP101	TRP62	ARG61	TRP63	ALA107	ARG112	ASP48	ASN59	ASN103	ASN106		
PaCS-MD$^{100,0.1}$1	ARG73	TRP62	TRP63	ASN59	ALA107								
	TRP62	ARG73	TRP63	ASN59	ALA107								
PaCS-MD$^{100,0.1}$2	ARG61	TRP62	TRP63	ARG73	ASN59	ALA107	ASP101	ASN103	ASN106				
PaCS-MD$^{100,0.1}$3	ASP48	ARG61	TRP62	TRP63	ALA107	ARG112	ASN103	ASN59	ASP101	ARG73	ASN106		
	ARG73	TRP62	TRP63	ALA107	ARG61	ARG112	ASN103	ASN59	ASP101	ASN106			
PaCS-MD$^{100,0.1}$4	ARG73	TRP62	TRP63	ALA107	ASN103	ASN59	ARG112	ASP101	ASN106				
PaCS-MD$^{100,0.1}$5	ARG73	TRP62	TRP63	ASP101	ASN59	ALA107	ASN103	ASN106					
PaCS-MD$^{100,0.1}$6	ARG73	TRP62	TRP63	ASN59	ALA107	ASP101	ASN103	ASN106					
PaCS-MD$^{100,0.1}$7	ARG61	TRP62	ASP48	ASN46	TRP63	ASN59	ALA107	ARG73	ASP101	ASN103	ASN106	GLY102	
PaCS-MD$^{100,0.1}$8	TRP62	ARG73	ASN103	TRP63	ASP101	ASN59	ALA107						
PaCS-MD$^{100,0.1}$9	TRP63	ARG73	TRP62	ASP101	ASN59	ALA107	ASN103	GLY102					
PaCS-MD$^{100,0.1}$10	ARG61	TRP62	TRP63	ASN59	ARG73	ALA107	ASP101	ASN103	ASN106				

Table 5. Residual order of hydrogen bond breakage with triNAG in SMD.

Trajectory	Hydrogen bonds of triNAG broken order												
	1	2	3	4	5	6	7	8	9	10	11	12	13
SMD^{fast}1	ARG61	ASP48	TRP62	TRP63	ASN59	ALA107	ARG73	ASP101					
SMD^{fast}2	ARG73	TRP62	TRP63	ASN59	ALA107	ASP101	ARG112						
SMD^{fast}3	ARG61	TRP62	ASP48	TRP63	ALA107	ASN59	ASP101						
SMD^{fast}4	ARG61	TRP62	ASP48	TRP63	ASN59	ALA107	ARG73	ASP101	ARG112				
SMD^{fast}5	TRP63	TRP62	ASN59	ASP101	ALA107	ASN103	ARG73	ARG112					
SMD^{fast}6	ARG73	TRP62	TRP63	ASN59	ALA107	ARG61	ASN46						
SMD^{fast}7	ARG73	TRP62	TRP63	ASN59	ALA107								
SMD^{fast}8	ARG61	TRP62	TRP63	ASN103	ARG73	ASP101	ALA107	ASN59					
SMD^{fast}9	TRP62	TRP63	ALA107	ASN59	ASN103	ASN106	ASP101						
SMD^{fast}10	ARG61	TRP62	TRP63	ASN59	ALA107	ASP101	ASN103	GLY102					
SMD^{fast}11	ARG61	ASP48	TRP62	TRP63	ASN59	ASP101	ALA107						
SMD^{fast}12	ARG73	ARG61	TRP62	TRP63	ASN59	ALA107	ASP101						
SMD^{fast}13	ARG61	TRP62	TRP63	ASN59	ALA107	ASP101	ARG112						
SMD^{fast}14	TRP62	TRP63	ASP101	ASN59	ALA107	ASN103							
SMD^{fast}15	ARG61	ARG73	TRP62	TRP63	ASN59	ALA107	ASP101						
SMD^{fast}16	TRP62	TRP63	ASP101	ALA107	ASN59	ARG73	ASN103						
SMD^{fast}17	ARG73	ASP101	TRP63	TRP62	ASN103	ASN59	ALA107						
SMD^{fast}18	TRP62	ARG73	TRP63	ASN59	ALA107	ASN103	ARG112	ASP101					
SMD^{fast}19	ARG61	ARG73	TRP62	TRP63	ASN59	ALA107	ASP101						
SMD^{fast}20	ARG61	TRP63	ASN59	ALA107	TRP62	ASP101	ARG73	ASN103					
SMD^{fast}21	ARG61	ASP48	TRP62	TRP63	ARG73	ASN59	ASP101	ALA107	ASN103				
SMD^{fast}22	ARG61	TRP63	ASP101	ALA107	ASN59	ARG73	TRP62	ASN103	GLY102	ASN106			
SMD^{fast}23	ARG61	TRP62	ARG73	TRP63	ASP101	ALA107	ASN59						
SMD^{fast}24	ARG61	ARG73	TRP62	TRP63	ALA107	ASN59	ASP101	ASN103	ARG112	ASN106	GLY102		
SMD^{med}1	ASP48	TRP62	ARG61	TRP63	ALA107	ASP101	ARG73	ASN59	ASN103	ARG112			
SMD^{med}2	ARG73	TRP62	TRP63	ASN59	ALA107	ASP101	ARG112						
SMD^{med}3	ARG61	ASP48	ARG73	TRP62	TRP63	ASN59	ALA107	ASP101					
SMD^{med}4	ASP48	ARG73	TRP62	ARG61	TRP63	ASN59	ALA107	ARG112	ASP101	ASN103	GLY102		
SMD^{med}5	ARG61	TRP62	ASP48	TRP63	ARG73	ASN59	ALA107	ASP101	ARG112				
SMD^{med}6	ARG61	ASP48	TRP62	TRP63	ARG73	ASN59	ASP101	ALA107	ASN103				

SMD^{med}7	ARG73	TRP62	ARG61	TRP63	ALA107	ASN59	ASP101	ARG112	ASN103				
SMD^{med}8	ARG61	ARG73	TRP62	TRP63	ASN59	ALA107	ASP101	ARG112	ASN103				
SMD^{med}9	ASP48	TRP63	ASN59	ALA107	TRP62	ARG73	ARG112	ASP101	ASN46	ASN103	GLY102	ARG61	ASN106
SMD^{med}10	ASP48	ARG61	TRP63	TRP62	ASN59	ALA107	ARG73	ASP101	ASN103	ARG112			
SMD^{med}11	ARG61	ASP48	TRP62	TRP63	ALA107	ASN59	ARG73	ASN103	ASP101				
SMD^{med}12	ARG73	TRP62	TRP63	ASN59	ALA107	ASP101	ASN103	ASN106	GLY102				
SMD^{slow}1	TRP62	TRP63	ASP101	ALA107	ASN59	ASN103	ARG73	GLY102	ASN46	ASN106			
SMD^{slow}2	ARG61	ARG73	TRP62	ASP48	ASP101	TRP63	ASN59	ALA107	ASN103	ARG112	ASN106	ASN46	
SMD^{slow}3	ARG61	TRP62	ASP48	ASN106	ARG112	ASN103	ASP101	TRP63	ASN59	ALA107	ARG73	ASN46	
SMD^{slow}4	ARG61	ARG73	TRP62	ASP48	ASN106	ARG112	ASP101	ASN103	GLY102	TRP63	ALA107	ASN59	ASN46
SMD^{slow}5	ARG73	ARG61	ASP48	ASN46	TRP62	ARG112	ASN103	TRP63	ALA107	ASN59	ASP101	GLY102	ASN106
SMD^{slow}6	ARG61	ASP48	ARG73	ASN46	TRP62	TRP63	ASN59	ALA107	ASN103	ARG112	ASP101	GLY102	ASN106
SMD^{slow}7	ARG73	ARG61	ASP48	TRP62	ASN106	TRP63	ALA107	ASN59	ASP101	ARG112	ASN103	ASN46	GLY102
SMD^{slow}8	GLY102	ASN103	ARG73	TRP63	ARG61	TRP62	ASP48	ASN59	ALA107	ASN46	ASP101	ASN106	ARG112
SMD^{slow}9	ASP48	ARG61	ARG73	TRP63	TRP62	ALA107	ARG112	ASN46	ASN59	ASN103	ASP101	ASN106	GLY102
SMD^{slow}10	ASP48	ARG61	ARG73	ASN103	TRP62	TRP63	ASN59	ALA107	ASP101	ASN46	ARG112		
SMD^{slow}11	ARG61	ASP48	ASN46	TRP62	ARG73	ARG112	TRP63	ALA107	ASN59	ASP101	ASN103	GLY102	
SMD^{slow}12	ARG61	ASP48	TRP62	ARG73	ASN46	ARG112	ASN59	TRP63	ALA107	ASP101	ASN103	ASN106	GLY102

4. Conclusion

PaCS-MD is a straightforward simulation algorithm, as is its implementation, and is nodoubt being suitable for parallel and/or distributed computing. In this work, we first showed that PaCS-MD easily induced the dissociation of LYZ and triNAG within the order of 10^0 – 10^1 *ns* and total cost of 10^2 *ns* MD simulation time (Table 1 and Fig 3). In contrast, no dissociation was observed during 1 μs of conventional MD (Fig 4). The cycles of multiple short MD simulations and the selection of rare events clearly accelerated dissociation. Although no bias was applied during MD, the dissociation speed of triNAG in PaCS-MD^{10,1} was equivalent to the pulling velocity of SMD^{fast}, whereas those in PaCS-MD$^{10,0.1}$ and PaCS-MD$^{100,0.1}$ were 5 times faster.

We examined the effects of the number of replicas (n_{rep}), MD length (t_{cyc}), velocity re-randomization, and snapshot selection on PaCS-MD sampling. As clearly shown by comparison of the results obtained using PaCS-MD$^{100,0.1}$ and PaCS-MD^{10,1} (Table 1 and Fig 5), increasing n_{rep} is a more efficient way to reduce the number of cycles necessary for dissociation than increasing t_{cyc} with a given fixed total simulation time per cycle, because the probability of observing rare events is proportional to n_{rep}. We found that velocity re-randomization enhanced sampling toward dissociation in the bound state, in which triNAG was trapped in energy minima by interaction with LYZ. In this situation, velocity re-randomization acted as a perturbation to enhance the occurrence of fluctuations toward escape from the bound state and selection raised the

probability of a rare event occurrence. Diffusion plays a more important role in the unbound state. Diffusion-governed movements occurring near the end of each MD run tend to be selected in PaCS-MD, which accelerates the dissociation process.

One trial of PaCS-MD generates a dissociation pathway as a combination of multiple short MD trajectories that mutually overlap in conformational space. We confirmed that the generated trajectories can be properly utilized to construct a MSM. The PaCS-MD-generated pathway is not a true dissociation pathway as a function of time, yet it can provide statistical information regarding the dissociation process along a zigzag-rod-like conformational space that connects the bound state to the completely unbound state. Interestingly, a drastic (10-fold) increase of n_{rep} in PaCS-MD$^{100,0.1}$ compared to PaCS-MD$^{10,0.1}$ only resulted in a slight increase in $\overline{\sigma^{each}}(d)$(Fig 10), indicating that PaCS-MD$^{100,0.1}$ sampled a conformational space similar to that in PaCS-MD$^{10,0.1}$ but more densely. This is related to the fact that we could build a MSM from each trial of PaCS-MD$^{100,0.1}$, whereas the statistics were insufficient in PaCS-MD$^{10,0.1}$. The longer (10-fold) simulation time, $\mathrm{t_{cyc}}$, in PaCS-MD^{10,1} generated an obviously thicker rod per trial, as shown by the value of $\overline{\sigma^{each}}(d)$ more than doubling. Since PaCS-MD$^{100,0.1}$ and PaCS-MD^{10,1} required the same computational time, PaCS-MD$^{100,0.1}$ densely sampled a narrower conformational space along the pathway, while PaCS-MD^{10,1} explored a wider space more sparsely. PaCS-MD$^{10,0.1}$/MSM resulted in a dissociation free energy ΔG_d slightly closer to the experimental value but the effect was relatively small. As long as sufficient statistics are achieved by sampling along the pathway, the width of the pathway, which shows the range of sampling almost perpendicular to the pathway, should not significantly affect ΔG_d. Although ΔG_d is in principle independent of the pathway, the actual value obtained by each trial likely contains some errors. Therefore, ΔG_d was calculated as the average over multiple pathways. Note that the average in the Jarzynski equality is taken over much thinner rods (i.e., SMD trajectories) which do not have spatial thickness in each trajectory. In this case, greater statistical accuracy can be achieved only by obtaining more trajectories. To obtain a more accurate ΔG_d, averaging over multiple trials may be more important than generating thicker rods.

As noted above, each trial may not provide adequate statistics for MSM. One approach to addressing this might be to gather trajectories from multiple trials and build one MSM. We constructed MSMs for all trajectories of the ten trials conducted using PaCS-MD[10,0.1] and obtained $\Delta G = 27.4(kJ/mol)$, consistent with experimental data. However, it should be noted that snapshots with similar d values can be assigned to one microstate, but snapshots from different trials can be very far in 3D space, especially in the unbound state, and thus should not be categorized in the same microstate. We judged that using d values to define microstates is invalid from the physical point of view.

As shown in Table 3, ΔG_d calculation with PaCS-MD/MSM yielded better results than calculation with PaCS-MD/US, SMD/US, or SMD/Jarzynski, as mentioned above. In addition, two different PaCS-MD/MSMs gave similar ΔG_d results. PaCS-MD/MSM showed the additional advantage that it gave the lowest standard deviation, which indicates the least free energy variation among the trials. Notably, the total simulation time for PaCS-MD/MSM is less than PaCS-MD/US or SMD/US because US requires longer simulation time (Table 1). The ΔG_dvalues from SMD/US and SMD/Jarzynski clearly depended on the SMD pulling velocity. In particular, SMD/Jarzynski overestimated ΔG_d in all cases. Detailed analysis of the LYZ-triNAG interactions in SMD showed that the dissociation order of the three NAGs from LYZ was different from that in PaCS-MD, suggesting that an unnatural dissociation process was induced by the force bias in SMD. The results of SMD/US indicated that artifacts caused by this bias were mostly recovered during US but velocity dependence was still observed. Therefore, the choice of pulling velocity is important in SMD/US simulations.

Chapter 4. Dissociation Peptide from Its Complex with Protein

1. Introduction

Free energy computation has long historical development throughout years. The question raised by Gumbart *et al.* on what the best strategy for binding free energies calculation stands still until now [24] since many new methods and improvements of the predecessors are still in the process of development. To overcome computational insufficiency in sampling, as a "cheap" end of free energy calculation methods, Molecular Mechanics Poisson-Boltzmann Surface Area (MM/PBSA)[77], Molecular Mechanics Generalized Born Surface Area (MM/GBSA)[78] have been developed and being widely used nowadays, despite their limitations including error in estimation the solute entropy, solvation free energies of charged, buried group, sampling and parameter tuning[79]. In the MM/PB-SA, the binding free energy is decomposed into several terms such as binding enthalpy, solvation free energy of binding and binding entropy[80]. Binding enthalpy can be determined from the simulations of the bound and unbound states of solutes, while solvation free energy of binding is derived from Poisson-based solvation model [81] and binding entropy usually is approximated by quasi-harmonic analysis [78]. Recently, Koehl *et al.* propsed a new version of MM/PB-SA, in which the Poisson-Boltzmann equations are modified to improve behavior in highly charged surface and the effects of the sizes of ion [82]. Because of its fast speed to estimate the binding free energy, MM/PB-SA is recently used as a new scoring function for protein-peptide docking in the well-known HADDOCK docking package[83].

Besides MM/PB-SA, two rigorous and widely used methods to obtain binding free energy are free energy perturbation (FEP) and thermodynamic integration, being classified as alchemical transformation method [84,85]. These methods sample the entire transformation path leading to more computational demanding, but more accurate than MM/PB-SA due to the inclusion of the route to compute the equilibrium binding constant [86]. However, the initial version of FEP using MD simulation is only limited to the small and rigid ligands. Later, Wou and Roux proposed a method [86] to calculate the binding free energy for protein and more flexible ligands using FEP technique [87,88] and US [89] with WHAM[90].Their idea is applying the various on-off-switchable restraining potentials to the flexible ligands in order to construct its PMF including the intermediate states. Therefore, the binding free energy can be expanded into multiple terms described

by the given the bias potentials. Their method successfully reproduced the binding energy of the human p56lck SH2 domain/AcpYEEI complex [86], and later T4 Lysozyme L99A Mutant in complex with Aromatic Molecules [25] and more cases not listed here. The methods later extended to protein-protein complex [26] whereas there are enormous degrees of freedom in the binding sites leading to the effort to design the restraining potentials. However, Gumbart *et al.* pointed out the limitation of this method that the enforcement of geometrical and conformational restraints would lead to the artificial reaction pathway which might give a rise to the free energy[23,26]. In addition, Yamashita and Fujitani pointed out that restraints of the protein structure by multi-step Target MD (mTMD) can eliminate the artificial deformation of the protein instead of SMD because the latter leads to the inaccurate estimation of PMF in US[29]. Those imply that if one can build the "more natural" dissociation pathways, the binding free energy calculation is more accurate.

Another important factor in binding free energy computation is method of sampling. Nowadays, there are more and more enhanced sampling techniques that can be named such as Replica Exchange MD[91], mTMD[29], Metadynamics[15], Parallel Cascade Selection MD (PaCS-MD) [31,47], Sparsity-weighted Outlier FLOODing [92]. Among of these, PaCS-MD is the enhanced sampling technique that does not apply bias forces to the system. In PaCS-MD, the state space is divided into multiple microstates, and the input configurations of parallel simulations are carried out by the recursive selection loop of the most toward the target final microstate. Therefore, the system is slowly moving from the initial microstate to the target microstate by restarting the multiple MD simulations from the selected configurations in the ranking process. Moreover, together with Markov State Model, PaCS-MD has been proved to be a powerful tool to estimate the protein/ligand binding free energy which allows all the flexibility of the ligands. Here in this work, we extend the proposed procedure to a protein/peptide case which essentially require tshe tolerance of flexibility in sampling method. We choose MDM2 protein bound to the transactivation domain of p53 peptide [93] as a test case. The transactivation domain of p53 peptide has been proved to be very flexible since it can adopt multiple conformations [94].

2. Calculation

We built two simulation boxes by using the crystal structure of the MDM2/TAD-p53 complex (PDB 1YCQ) after modeling the missing hydrogen atoms, capping with the acetylated C-terminal, neutral N-terminal and missing residues as shown in Fig 13. After that, the complex was solvated into a box of $18.8\times9.9\times8.9\ nm^3$ with TIP3P water molecules. Sodium and chloride ions were added to the simulation box to realize ion concentration of 0.15M and charge neutrality. We used AMBER99SB-ILDN for the protein complex. All simulations are performed using GROMACS 5.1.2[55].

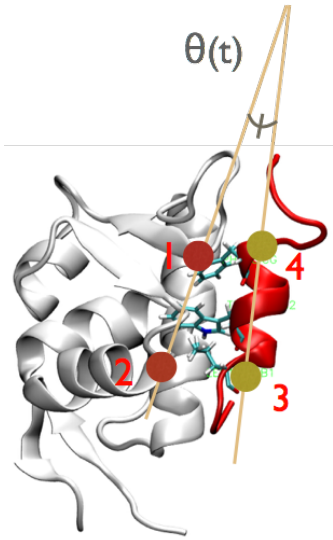


Fig 13. Structure of solvated MDM2/TAD-p53 complex after remodeled missing residues and equilibrium molecular dynamics simulation. TAD-p53 is colored in red. Point 1 and 2 depict for the centers of mass of the MDM2 upper part and lower part residues respectively, while point 3 and 4 depict for the centers of mass of TAD-p53 upper part and lower part residues respectively. The upper and lower parts is defined by the initial position of TAD-p53 residue TRP23.

The solvated models were then energy-minimized by the steepest descent method followed by the conjugate gradient method with $1000\ kJ/mol.nm^2$ heavy atom positional restraints to keep the crystal contacts between binding interface of MDM2/TAD-p53. The equilibration MD simulations were carried out the same as the procedure written in chapter 3, section 2 of this dissertation except in NVT ensemble simulation, the system was heated up from 0 K to 400 K within 1 ns and equilibrated at 400 K for 500 ps, then cooled down to 300 K in next 500 ps and equilibrated at 300 K for 1 ns. In addition, we only applied the position restrained to the non-modeled region

of the complex (residues that is in PDB structure). This exception NVT ensemble simulation ensures the configuration of modeled missing residues to be in lower energy configuration. After that, we carried out PaCS-MD simulation of dissociation. We used 30 replicas and 0.1-ps short MD simulations for both systems. We performed 10 ns MD simulation for each system and pick up 10 different configurations from the former simulation as the initial snapshots for the next cycle.

3. Result and discussion

3.1. Equilibration of the remodeled system MDM2/TAD-p53

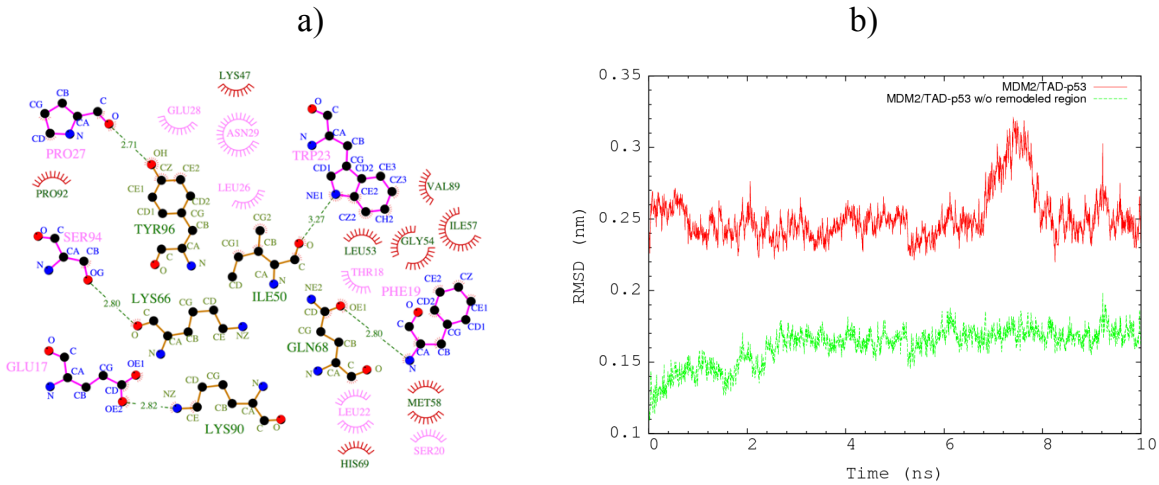


Fig 14. Maintaining of the key interaction between MDM2 and TAD-p53. a) Key interaction in the binding interface. Pink label denotes for TAD-p53 while green label denotes for MDM2. Green dash lines mark for hydrogen bonds and their number show distances of hydrogen bonds in Å unit. b) Root mean square deviation of MDM2/TAD-p53 complex after least-squared fitting with MDM2 crystal structure.

The MDM2 and TAD-p53 complex in the bound form have 5 hydrogen bonds, 17 hydrophobic interactions in the binding interface as shown in Fig 14a). The binding interface between TAD-p53 and MDM2 consists of one helix and two turns of MDM2, and one helix of TAD-p53 as in Fig 13. The equilibration simulation maintained the key interactions of the complex the same as in the crystal structure, i.e., hydrogen bonds between ILE50-TRP23, GLN68-PHE19, VAL89-LEU22, TYR96-PRO27 (note that the former residues belong to MDM2 and the latters belong to TAD-p53). Two key

amino acids, TRP23 and PHE19, stayed inside the hydrophobic cleft of MDM2. The dominant of hydrophobic interactions indicate that the weak but specific binding of the complex was maintained as indicated by S-W. Chi et al [95]. The helical content in TAD-p53 peptide span from PHE19 to LEU25 by hydrogen bond network was the same as found in the work by Joseph *et al.*[96]. In addition, by performing computational mutagenesis, Moreira *et al.* found that MET20 and TYR22 are also suggested to be new hotspots, besides the hotspot residue PHE19, TRP23 and LEU26[97]. Moreover, they indicated that p53-backbone hydrogen bonds with MDM2 are important.

To examine whether the protein complex is stable, we calculated the RMSD of the MDM2/TAD-p53 complex after performing least-squared fitting with MDM2 crystal structure as shown in Fig 14b). The region included in the crystal structure was stable while the overall protein structure was slightly shifted in RMSD from 6 to 8 ns of the equilibration simulation. However, the overall change in RMSD is less than 0.3 nm indicating the stability of the system.

3.2. Free energy difference of dissociation between MDM2/TAD-p53

Without any bias force, PaCS-MD easily simulated dissociation pathways. The dissociation of MDM2/TAD-p53 can be splitted into 3 steps as shown in

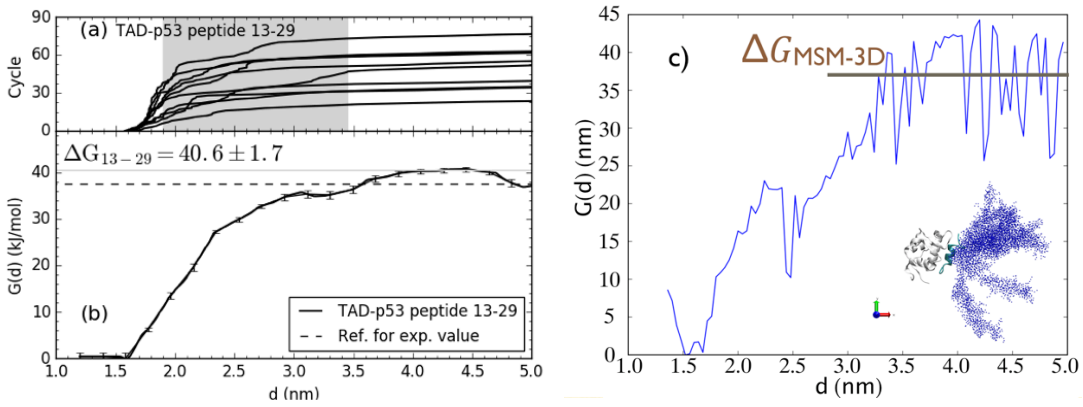


Fig 15a): first, TAD-p53 slightly increased the inter-COM distance in the bound state (d is lower than 1.9 nm), which induced the increment in the number of water molecules inside the binding interface, then a part of TAD-p53 achieve the dissociation from

MDM2 in the partially bound state, which is the grey highlight in

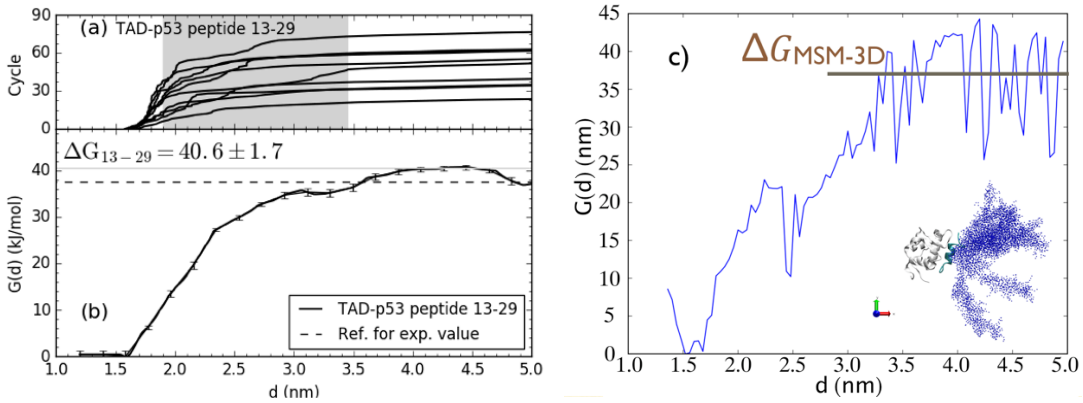
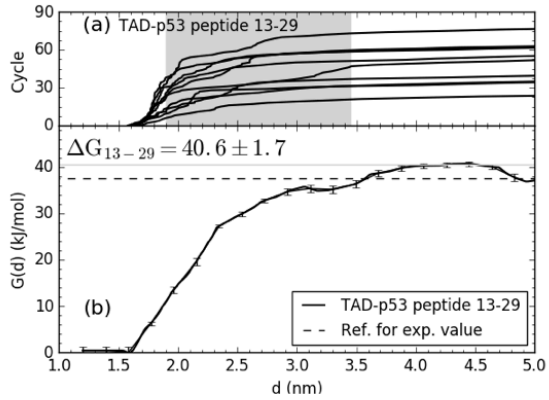


Fig 15a), and final complete dissociation with *d* larger than 3.4 nm in the unbound state.The linear increase of inter-COM distance was found in the bound state and unbound state in contrast to the partially bound state.

By performing PaCS-MD, we generated the dataset of inter-COM distance of all short simulations of all cycles. The Markov State Model is constructed from these datasets using the Maximum Likelihood Estimation procedure. After carefully considering the evolution of implied timescale with lagtime[44], we chose a lag time of 20 ps to build MSM in this work. From MSM, the population of microstate is calculated



and taken average to obtain the PMF.

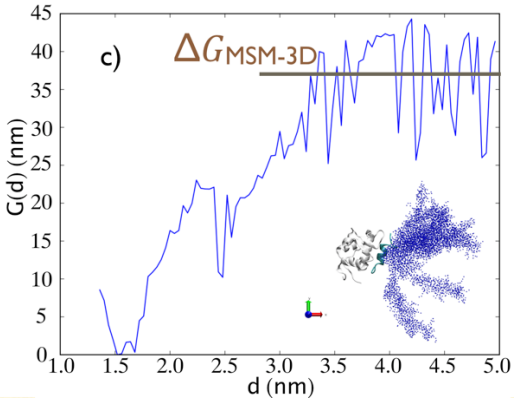
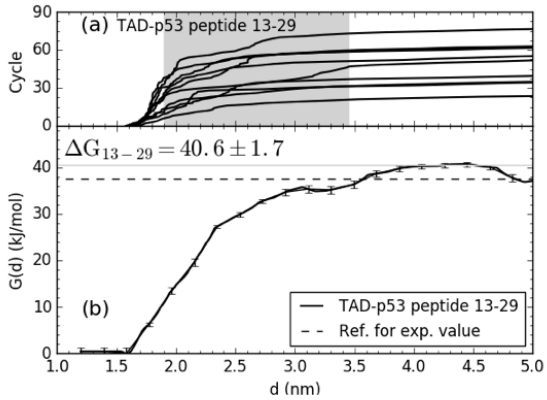


Fig 15b) shows the free energy of dissociation of our simulation. The free energy difference between the bound state and the plateau region, $4.0 \leq d \leq 4.5$ nm was considered as the dissociation free energy, which was $\Delta G_d^{all} = 40.6 \pm 1.7\ kJ/mol$, comparable to the experimental value of 37.7 kJ/mol[98]. There are two plateau regions in



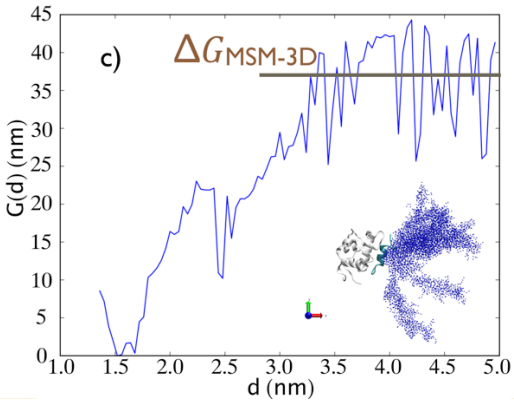


Fig 15b) i.e. from 3.0 to 3.4nm and 4.0 to 4.5nm. The former is mostly related to the partially bound state of p53 with the binding cleft of MDM2 where a few contacts

of TAD-p53 with MDM2 were left. The second plateau indicates the complete dissociation of the complex.

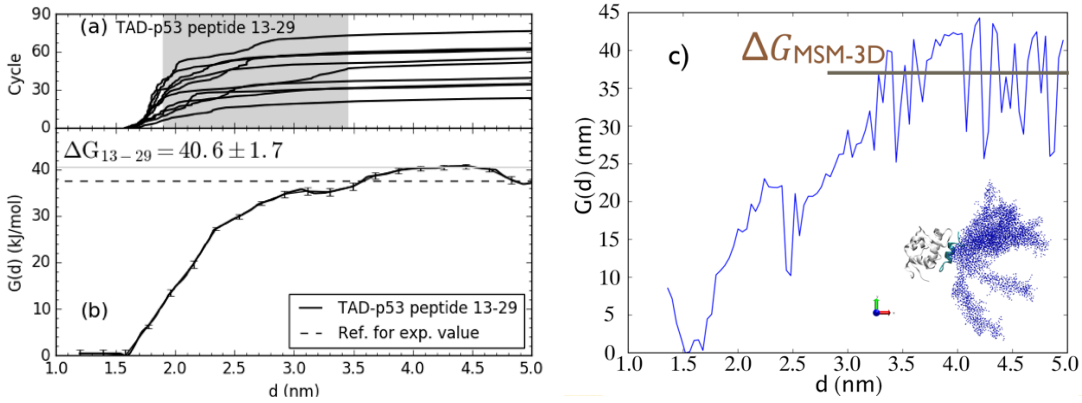


Fig 15. Evolution of a) inter-COM distance with cycle, b) free energy of dissociation with inter-COM distance by MSM using single-trial inter-COM distance dataset, c) free energy of dissociation with inter-COM distance by MSM using 3D dataset. Error bar in b) shows the calculation standard error. Grey highlight in a) show the partially bound state of MDM2/TAD-p53 complex. Inset of c) shows the representative positions of microstates in 3D.

To have more statistics in the dataset, we used the 3D coordinates of COM of TAD-p53 in all trajectories of all trials as a single large dataset. We performed clustering on the dataset by using Cartesian clustering which yields total number of 2751 microstates in our dataset (as shown in the inset of Fig 15c). After that, we build MSM based on this result using lagtime 25 ps to obtained the evolution of free energy on the TAD-p53 COM coordinates. After that, we convert from 3D space to inter-COM distance as shown in Fig 15c). The obtained binding free energy $\Delta G_{MSM-3D} = 37.2 \pm 6.3\ kJ/mol$, which is in good agreement of the experimental results. Furthermore, we extract Mean First Passage Time matrix and calculate the association rate constant via $k_{on} = 1/(MFPT_{on}C^{comp})$, whereas C^{comp} is the concentration regarding to the simulation box. The obtained association rate constant $k_{on} = 8.9\times10^{6}\ M^{-1}s^{-1}$, being in agreement with Schon et al.' work ($9.2\times10^{6}\ M^{-1}s^{-1}$)[99]. We conclude that we can both estimate binding free energy and rate constant via PaCS-MD/MSM accurately.

3.3. Structural changes during dissociation

To understand more about the dissociation pathways of TAD-p53 from MDM2, we calculated the binding interface angle fluctuation and RMSD of TAD-p53 during the

dissociation using the concatenated trajectories compared to the initial structure (Fig 13). The binding interface angle fluctuation $\Delta\theta(t)$is defined as followed.

$$\Delta\theta(t) = \theta(t) - \theta(0) \quad (17)$$

$$\theta(t) = 180^o - \left[cos^{-1}\left(\frac{d_{1,2}^2+d_{2,3}^2-d_{1,3}^2}{2d_{1,2}d_{2,3}}\right) + cos^{-1}\left(\frac{d_{2,3}^2+d_{3,4}^2-d_{2,4}^2}{2d_{2,3}d_{3,4}}\right)\right] \quad (18)$$

where $d_{x,y}$ is the distance between the center of mass of the group of atoms x and group of atoms y. Here we define the group of atoms as follows. Group 1 and group 2 are the atoms of the binding interface of MDM2 (all atoms in the same residues that the closest distances between pairs of atoms belonged to different group are less than hydrogen bonds distance), in which group 1 is the part have position y > 4.7 nm and group 2 is y ≤ 4.7 nm (). Group 3 and group 4 are the atoms of TAD-p53 that divided in the way the same as group 1 and group 2.

The lower $|\Delta\theta|$ indicates the direction of dissociation of TAD-p53 whether it is perpendicular with the initial structure or not. Fig 16a) shows there were 4 trials having $max|\Delta\theta|$ fluctuated around 0^o, which indicates the dissociation direction was mostly perpendicular to the binding interface. The $\pi - \pi$ stacking interactions were formed between PHE19-TRP23 of TAD-p53 and PHE19-TYR44 of TAD-p53 and MDM2 respectively. These two interactions mostly maintained the dissociation direction perpendicular to the binding interfaces. In this case, the main contribution to the RMSD change mainly occured from the modeled loops of TAD-p53. On the other hand, $\Delta\theta$ were significantly positive (Fig 16a). The dissociation only happened perpendicularly to the binding interface (pathway type 1) or the PHE19 dissociates before LEU25 (pathway type 2). We observe no case in which LEU25 dissociated before PHE19 did. This indicates the importance of PHE19, TRP23 and LEU25 of TAD-p53 which agrees with the other experimental and computational works[93,97].

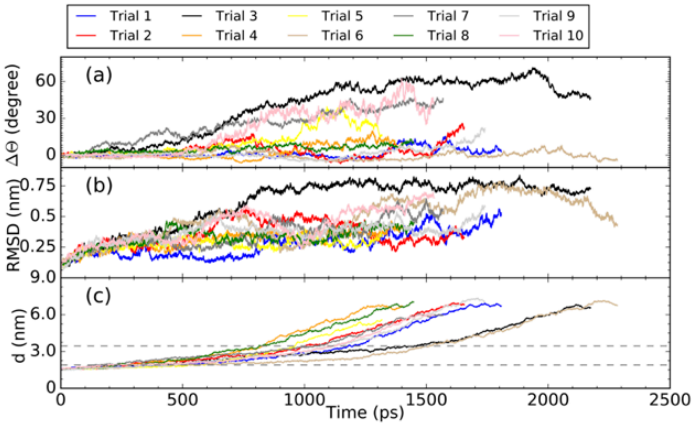


Fig 16. Structural changes of TAD-p53 during dissociation from the complex of MDM2. The concatenated trajectories' evolution of a) The binding interface angle fluctuation between MDM2 and TAD-p53, b) Root mean squared deviation of TAD-p53, and c) inter-COM distance of MDM2/TAD-p53 in reactive trajectories.

An interesting question can be raised: what is the most favorable dissociation pathway? To answer this question, we individually calculated the free energy difference of dissociation for each pathway (via the population of each single pathway represented in term of free energy) and analyze the results in considering the RMSD and inter-COM distance of TAD-p53 in reactive trajectories (Fig 16 b and c). We found that either dissociation pathway type 1 or type 2, if TAD-p53 performs the structural changing before entering to the partially bound-state (in which RMSD is larger than 0.3 nm), the ΔG_d is larger than the other cases. In spite of the same tendency of structural change, pathway type 2 (trial 3 has $\Delta G_d = 59.7\ kJ/mol$) has a larger ΔG_d than that in pathway type 1 (trial 8 has $\Delta G_d = 43.53\ kJ/mol$). Therefore, we conclude that the internal structural change more contributed to ΔG_d than the direction of the dissociation. Moreover, pathway type 1 has shorter dissociation time than pathway type 2. Note that the dissociation pathway that has the ΔG_d value closest to the experimental value was perpendicular to the binding interface and less structural change during the dissociation (lower RMSD value) than the rest.

4. Conclusion

In this thesis, we carried out the calculation of dissociation free energy of protein/peptide complex (MDM2/TAD-p53), which is strongly related to cancer. We successfully reproduce experimental value of dissociation free energy[98] and better than the estimation using MM/PBSA[100]. This confirmed the effectiveness of PaCS-MD with MSM in binding free energy calculation for the flexible peptide in complex with protein. In addition, we prove that via PaCS-MD/MSM, the rate constant can be estimated accurately.

Moreover, we analyzed the most favorable TAD-p53 pathway of dissociation from the complex with MDM2 in details. We found that the dissociation should contain structural change which are not suitable for biased simulation as discussed elsewhere above. Moreover, dissociation perpendicular to the binding interface occured with the least cost of structural change and free energy.

Chapter 5. Flexible Docking of Protein/Peptide Complex

1. Introduction

Molecular docking is considered as an essential tool in structural biology and the other related fields[101]. Up to the time of writing this dissertation, molecular docking methods are very diverse with more than 60 different docking tools, and can be divided into rigid-body docking and flexible docking[102]. Rigid body docking is based on the famous assumption "lock and key" of Fischer, i.e., the target and docked molecules can be treated as rigid bodies and docking finds the most suitable fit position of the complex[103]. However, later this assumption is revised into "induced-fit" theory of Koshland *et al.* which considers flexilibity of both molecules. However, up to now, most of the successful softwares in the CAPRI contest[104] still take advantage of the rigid-body docking to generate decoy structures as much as possible, refining the decoys and performing the ranking of decoys after that. The most famous approach to generate the decoys is to take advantage of the Fast Fourier Transformation by using interaction terms[105] or 3D space[106] as the inputs to explore the space of docked conformations. This approach is commonly used due to the cheaper cost of computation to generate decoys. These rigid-body docking methods do not work well for very flexible ligands such as small peptides.

In contrast to rigid-body docking, flexible docking considers all the movement of sidechain of ligand molecule with rigid-body treatment of receptor molecule, which can be classify into four types: a) simulation-based docking of complete molecules, b) *in-site* combinatorial search, c) ligand build up and d) site mapping and fragment assembly[102]. In simulation-based methods, computational costs are the most draw-back to explore whole free energy landscape of binding compared to the fast rigid-body docking methods. Therefore, enhanced sampling simulation technique is required for this scheme to be efficient. Examples of recently developed methods based on molecular dynamics simulation are CDOCKER[107], and MedusaDock[108,109]. CDOCKER is a molecular dynamics simulation based flexible docking method using simulated annealing with CHARMM force field and soft-core potential to perform multiple replicated simulations to search for decoys[107]. MedusaDock takes advantage of implicit solvent Medusa force field to improve the flexibilities of ligands and receptors upon binding[109]. However, by using implicit solvent force field, MedusaDock neglects the

roles of solvent molecules in the binding interfaces which often affect the binding conformation of the complex.

Here we consider that the keys of the successful MD simulation-based flexible docking methods are to take account of all the flexibility of both ligand and receptor, which treats the desolvation of the binding interface upon binding. PaCS-MD was proved to be a unbiased effective enhanced sampling technique that can accelerate the dissociation simulation. Moreover, during dissociation by PaCS-MD, the conformations of the ligand and receptor changed due to the interactions between them, which indicate that PaCS-MD is suitable for generating different conformations of ligand and receptor. Moreover, PaCS-MD can be incorporated with all-atom forcefield with high accuracy. Therefore, we are encouraged to carry out the flexible docking based on PaCS-MD.

2. Calculation

We generated the unbound simulation box of MDM2 and TAD-p53 in which the inter-COM distance between MDM2 and TAD-p53 larger than 5 nm. The configuration of TAD-p53 was taken from our dissociation simulation using PaCS-MD. Therefore, the initial structure of TAD-p53 is not the same structure as the complex structure in crystal (PDB ID: 1YCQ). MD simulation parameters used in this chapter are the same as those used in the previous chapter. We employed PaCS-MD scheme for association and dissociation consecutively to repeat the cycles of TAD-p53 association and dissociaiton. We used inter-COM distance for the selection quantity of PaCS-MD again. Our simulation first starts with association. Here we introduced the switching criteria between association and dissociation simultion. If the top one snapshot occured within the first 20% of the MD time, which is $t_{selected} < 0.2t_{cyc}$, PaCS-MD switches to dissociation, in which the top one inter-COM is denoted as d_{switch}. If $d > d_{switch}$ is true, PaCS-MD switches from dissociation to association again. By repeating switch between association and dissociation, we generatee multiple complex structures of MDM2/TAD-p53.

3. Result and discussion

3.1. Flexibility of TAD-p53

We aim to apply this method to systems including flexible peptide because this has not been well established yet. Therefore, we first examined flexibility of TAD-p53 to check whether it is suitable as our target. We explore the free energy landscape of TAD-p53 by well-tempered Metadynamics with two collective variables: radius of gyration and the distance between the N- and C-termini (N-C distance). The initial height and width of Gaussian bias potential are set to be 0.5 kJ/mol.nm and 0.01 nm respectively for both collective variables. Total simulation time for metadynamics is 100 ns. After that, we calculated the free energy landscape of TAD-p53 as shown in Fig 17.

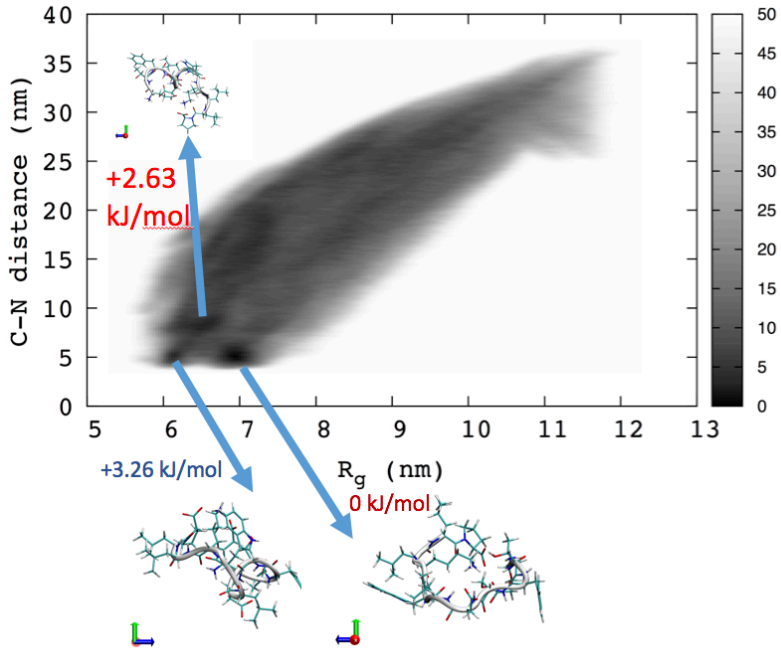
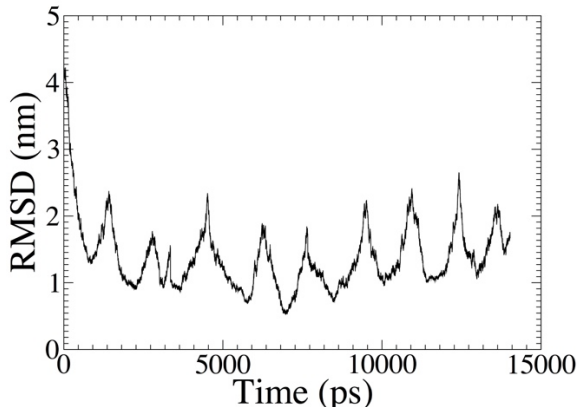


Fig 17. Free energy landscape of TAD-p53 constructed by the well-tempered metadynamics. The colormap unit is in free energy unit kJ/mol.

The lowest free energy conformation of TAD-p53 (denoted as 0 kJ/mol in Fig 17) does not have specific secondary structure. The global minima conformation has radius of gyration at 6.9 nm, and C-N distance is at 5 nm. Another local minima which have the same N-C distance is at 3.26 kJ/mol higher than the global minima conformation, which is at 6.1 nm radius of gyration. Interestingly, we find that the local minima being

near the bound conformation of TAD-p53 to MDM2 is at +2.63 kJ/mol compared to the global minima. This is in agreement with the other works related to MDM2/p53 complex that p53 is assumed to be low cell permeability and proteolytically unstable, binding-induced folding protein[95,110]. This can be explained by the fact that interactions between MDM2 and TAD-p53 are mostly hydrophobic interactions. However, the special motif of 3 key interactions that discussed in part 2 of this chapter is unique for MDM2/TAD-p53 complex. In addition, free energy difference between the local-local minima, local-global minima are not high, less than 4.184 kJ/mol that makes TAD-p53 is flexible peptide. Therefore, we confirm the choice of TAD-p53 and MDM2 is suitable for our purpose.

3.2. Generating the bound conformations



1.

Fig 18. Evolution of RMSD of TAD-p53 with the crystal structure PDB id 1YCQ. RMSD is calculated after fitting MDM2 backbone with its crystal structure.

As explained in section 3.2, we performed cycles of association and dissociation PaCS-MD up to 274 cycles by changing the ranking of inter-COM distance as shown in Fig 18, which shows the RMSD of TAD-p53 of a reactive trajectory after performing least square fitting with the backbone of MDM2 in crystal structure (PDB id 1YCQ). In each cycle of association and dissociation, the RMSD values of local minima were different, and the bound conformations of the MDM2/TAD-p53 complex were also significantly different. To examine the diversity of the bound conformations, we plotted all the inter-COM distance of all snapshots with the RMSD from the crystal structure of TAD-p53 during cycle in Fig 19, which clearly indicates that the inter-COM distance and RMSD

from the crystal structure spanned large conformational space within the proximity of cycle. In the bound state where inter-COM distance lower than 1.9 nm, the highest RMSD was ~ 1.5 nm, which indicates the significant difference among bound structures. In addition, the lowest inter-COM distance does not correspond to the lowest RMSD conformation and vice versa. In fact, this is true because the distance from the COM of MDM2 to its surface is not uniformly distributed in all direction.Therefore, one may ask whether inter-COM distance is the good reaction coordinate (or monitoring variable) of docking PaCS-MD. The advantage of using inter-COM distance is that it is not specific for any points on the surface of target protein, which does not lead to loss of generality. In addition, inter-COM distance here played the role of determining the choice of association or dissociation. The movement of ligand (docked peptide) toward the target protein is governed by diffusion and interaction (as discussed in part 1 of this chapter), which is mostly affected by the long range interaction and ionic strength of the solution. Fig 19 also implies that structures of TAD-p53 sampled here are similar with those in well-tempered metadynamics as discussed in chapter 3.

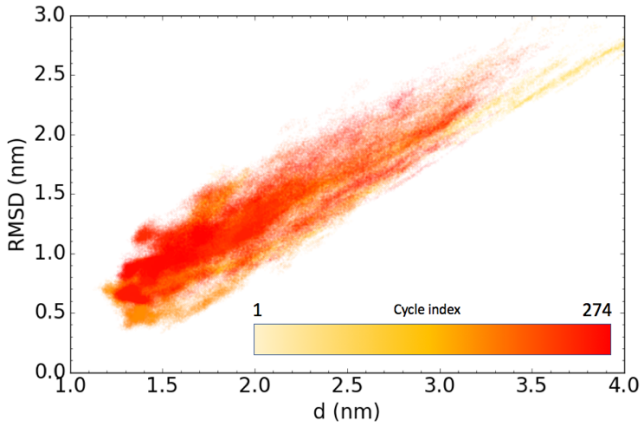


Fig 19. Inter-COM distance of MDM2/TAD-p53 complex and RMSD with crystal structure of TAD-p53 after least square fitting of MDM2. White area indicates the non-data area.

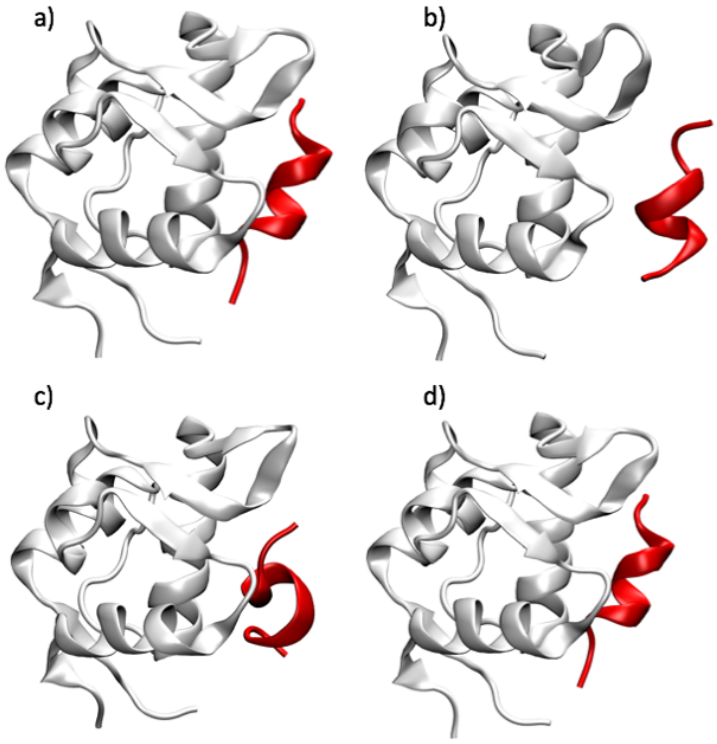


Fig 20. Representative structure of the bound conformation clusters obtained from PaCS-MD. a-d denote for the cluster 1-4. Note that a

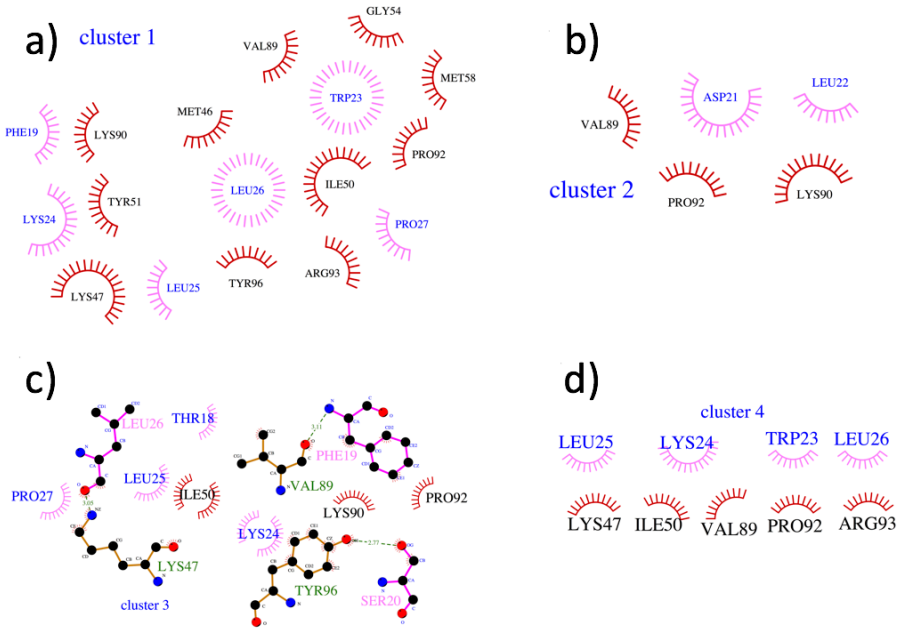


Fig 21. Interaction analysis of the representative structure of the bound conformation clusters of MDM2/TAD-p53. Cluster number to the order of Fig. 20

We performed clustering of all the bound state structures and obtained 4 clusters. The representative structures of each cluster are shown in Fig 20, and detailed analysis of interactions is shown in Fig 21. Among the four structures shown in Fig 20, structure a) and d) which to cluster 1 and cluster 4 respectively, are similar with the native structure of MDM2/TAD-p53 complex due to the significant large helix content in TAD-p53. Interestingly, although the unbound conformation of TAD-p53 does not directly face to the binding cleft of MDM2, all the representative structures bound with the correct binding interface. Cluster 1 contained the structure with the lowest RMSD from the crystal structure (RMSD = 0.429 nm). The main differences between the best generated structure and crystal one occurred in the side chain of PHE19 and TRP23, which should be inside the binding cleft and form the $\pi - \pi$ stacking interactions. However, if we only calculate the RMSD of the binding sites, the RMSD decreased to 0.243 nm, which indicates that the best generated structure properly reproduced interactions at interface. Cluster 1 contained all the hydrophobic interactions except for 3 hydrogen bonds in cluster 3 (Fig 21). Although the same backbone structure was formed as in cluster 1, cluster 4 sidechain is 180 degree rotated about the backbone axis from cluster 1 and the key residues PHE19 and TRP23 pointed out of the binding interface. The hydrophobic interactions between MDM2/TAD-p53, therefore, is replaced by the opposite side of TAD-p53, rather than the side of PHE19 in cluster 4.

Our preliminary results of docking simulation is promising because TAD-p53 can find the correct binding pose and good orientation similar to the crystal structure. As a blind docking, we need to score the bound structures. To determine the most probable structure, we will use MSM as shown below.

3.3. Predicting the best complex structure via MSM

The disadvantage of most docking methods is that they use scoring function that does not reproduce actual binding free energy. In fact, simplified representaion cannot include the high resolution information of the inter-molecular interactions. One possible solution is to perform separate binding free energy calculation by fast methods to rank the generated decoys. The more accuracy you gain, the more computing resource you need. Here we propose to rank the decoys base on the MSM-derived free energy

difference. We used the free energy difference obtained from MSM of all the trajectories to rank the obtained conformation in the bound state. We build the dataset of dRMSD from the initial input conformation of TAD-p53 after fitting to monomer conformaiton of MDM2. The PMF is shown in Fig 22. The minima of the PMF is at 4.21 nm. Interestingly, the lowest RMSD structure that we showed in part 3.3.2 is included in the microstate of the PMF global minimum.

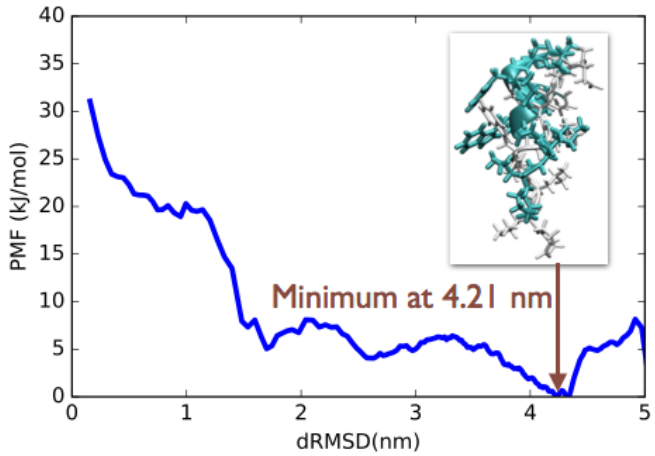


Fig 22. Evolution of PMF with the distance RMSD of TAD-p53 after least square fitting to MDM2. The inset shows the representative structure at the minimum of PMF

Here, we discuss the efficiency of our proposed docking methods. First of all, we expect that most of flexible protein/peptide docking can be treated by this method. Second, the PMF obtained from the PaCS-MD trajectories and MSM takes into account of the solvation and desolvation of binding interface. The scoring function in our method is calculated as the binding free energy with all-atom model with explit solvent. Third, we use enhanced sampling technique that is faster to generate the bound conformation than other traditional classical simulation. Moreover, the PaCS-MD method is highly suitable for distributed computing system, which accelerates the sampling. However, our current method has not reproduce complete desolvation of the binding interface upon binding. This is probabliy due to slow timescale of dehydration, which was not accerelated by PaCS-MD. We are now working to solve this problem.

4. Conclusion

Here we introduced a method of flexible docking based on PaCS-MD. Note that our methods allows mutual adaptive conformational changes of both protein and peptide upon interaction. Through our first preliminary test case of MDM2/TAD-p53, the proposed method was successful in finding the correct binding interface of MDM2 without any prior-knowledge of crystal structure. We consecutively performed cycles of association and dissociation simulations to generate the bound conformations, and ranking with binding free energy derived from MSM overcame the scoring problem in other docking methods.

Chapter 6. Concluding remarks

In this dissertation, we thoroughly examined two features that we pointed out: 1) whether the calculated binding free energy reproduces experimental value if more natural pathways by unbiased MD; 2) Whether we can calculate the binding free energy directly from trajectories generated by PaCS-MD. By using PaCS-MD and MSM, we have proved that we can yield the better free energy differences than the other methods, SMD/US in the case of protein/ligand and MM/PB-SA in the case of protein peptide. Moreover, we investigated acceleration mechanism of PaCS-MD in chapter 3 which had not been quantified, and then proposed a flexible docking method for a flexible protein-peptide complex. Since recent increase of computional power has been being achieved by the increase of the number of parallel processors rathen than the speedup of each processors, PaCS-MD is a promising tool for computational biophysics, computational drug design due to its high compatibility with parallel and distributed computings.

In addition, in this dissertation, we examined the dissociation pathways generated by SMD and PaCS-MD in detail. PaCS-MD tends to provide more natural pathways than SMD does. We also discussed the conditions to achieve more efficient samplig with PaCS-MD: the increase in the number of replicas is better than the increse in the trajectory length. It is worthwhile noting that PaCS-MD/MSM consumes less computing resources than SMD/US with the same scale of accuracy.

Moreover, we also performed detailed analysis on the dissociation of TAD-p53 from the complex with MDM2. TAD-p53 is strongly related to cancer. We found that the most favorable dissociation pathway occurs with less conformational change and dissociation perpendicular to the binding interface.

Our next plan in the near future is to improve our docking method so that it has ability to predict a correct conformation which agrees with crystal structures. Another direction is to develop a method in which PaCS-MD simulation method is combined with machine learning methods to improve the selection procedure.

References

1. Carpenter, B., Nehmé, R., Warne, T., Leslie, A. G. W. & Tate, C. G. Structure of the adenosine A2A receptor bound to an engineered G protein. *Nature* 1–15 (2016). doi:10.1038/nature18966
2. Katritch, V., Cherezov, V. & Stevens, R. C. Structure-function of the G protein-coupled receptor superfamily. *Annu. Rev. Pharmacol. Toxicol.* **53,** 531–56 (2013).
3. Chen, Y.-C., Huang, S.-H. & Wang, S.-M. Adenosine-stimulated adrenal steroidogenesis involves the adenosine A2A and A2B receptors and the Janus kinase 2–mitogen-activated protein kinase kinase–extracellular signal-regulated kinase signaling pathway. *Int. J. Biochem. Cell Biol.* **40,** 2815–2825 (2008).
4. Okonkwo, D. O. *et al.* A comparison of adenosine A $_{2A}$ agonism and methylprednisolone in attenuating neuronal damage and improving functional outcome after experimental traumatic spinal cord injury in rabbits. *J. Neurosurg. Spine* **4,** 64–70 (2006).
5. Velazquez-Campoy, A. *et al.* in *Current Protocols in Cell Biology* 17.8.1-17.8.24 (John Wiley & Sons, Inc., 2004). doi:10.1002/0471143030.cb1708s23
6. Schuck, P. USE OF SURFACE PLASMON RESONANCE TO PROBE THE EQUILIBRIUM AND DYNAMIC ASPECTS OF INTERACTIONS BETWEEN BIOLOGICAL MACROMOLECULES. *Annu. Rev. Biophys. Biomol. Struct.* **26,** 541–566 (1997).
7. Gauthier, T. D., Shane, E. C., Guerin, W. F., Seitz, W. R. & Grant, C. L. Fluorescence quenching method for determining equilibrium constants for polycyclic aromatic hydrocarbons binding to dissolved humic materials. *Environ. Sci. Technol.* **20,** 1162–1166 (1986).
8. Hulme, E. C. & Trevethick, M. A. Ligand binding assays at equilibrium: validation and interpretation. *Br. J. Pharmacol.* **161,** 1219–1237 (2010).
9. Pollard, T. D. A guide to simple and informative binding assays. *Mol. Biol. Cell* **21,** 4061–7 (2010).
10. Lelièvre, T., Rousset, M. & Stoltz, G. *Free Energy Computations*. (Imperial College Press, 2010). doi:10.1615/AtoZ.f.free_energy
11. Kirkwood, J. G. Statistical mechanics of fluid mixtures. *J. Chem. Phys.* **3,** 300–313 (1935).
12. Zwanzig, R. Nonlinear generalized Langevin equations. *J. Stat. Phys.* **9,** 215–220

(1973).

13. 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).
14. Jarzynski, C. Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.* **78,** 2690–2693 (1997).
15. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 12562–12566 (2002).
16. Wang, F. & Landau, D. P. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* **86,** 2050–2053 (2001).
17. Christ, C. D., Mark, A. E. & van Gunsteren, W. F. Basic ingredients of free energy calculations: A review. *J. Comput. Chem.* **31,** NA-NA (2009).
18. Hornak, V. *et al.* Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function and Genetics* **65,** 712–725 (2006).
19. Oostenbrink, C., Villa, A., Mark, A. E. & Van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **25,** 1656–1676 (2004).
20. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Meth* **14,** 71–73 (2017).
21. William L. Jorgensen, *, David S. Maxwell, and & Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. (1996). doi:10.1021/JA9621760
22. Ramírez, C. L., Martí, M. A. & Roitberg, A. E. in *Methods in Enzymology* **578,** 123–143 (2016).
23. Deng, Y. & Roux, B. Computations of standard binding free energies with molecular dynamics simulations. *J. Phys. Chem. B* **113,** 2234–2246 (2009).
24. Gumbart, J. C., Roux, B. & Chipot, C. Standard Binding Free Energies from Computer Simulations: What Is the Best Strategy? *J. Chem. Theory Comput.* **9,** 794–802 (2013).
25. Deng, Y. & Roux, B. Calculation of standard binding free energies: Aromatic molecules in the T4 lysozyme L99A mutant. *J. Chem. Theory Comput.* **2,** 1255–1273 (2006).
26. Gumbart, J. C., Roux, B. & Chipot, C. Efficient Determination of Protein–Protein Standard Binding Free Energies from First Principles. *J. Chem. Theory Comput.* **9,**

3789–3798 (2013).

27. Sugita, Y., Kitao, A. & Okamoto, Y. Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.* **113,** 6042–6051 (2000).
28. Schlitter, J., Engels, M. & Krüger, P. Targeted molecular dynamics: A new approach for searching pathways of conformational transitions. *J. Mol. Graph.* **12,** 84–89 (1994).
29. Yamashita, T. & Fujitani, H. On accurate calculation of the potential of mean force between antigen and antibody: A case of the HyHEL-10-hen egg white lysozyme system. *Chem. Phys. Lett.* **609,** 50–53 (2014).
30. Zwier, M. C. & Chong, L. T. Reaching biological timescales with all-atom molecular dynamics simulations. *Curr. Opin. Pharmacol.* **10,** 745–752 (2010).
31. Harada, R. & Kitao, A. Parallel cascade selection molecular dynamics (PaCS-MD) to generate conformational transition pathway. *J. Chem. Phys.* **139,** (2013).
32. Harada, R. & Kitao, A. Nontargeted Parallel Cascade Selection Molecular Dynamics for Enhancing the Conformational Sampling of Proteins. *J. Chem. Theory Comput.* **11,** 5493–5502 (2015).
33. Peng, J. & Zhang, Z. Unraveling low-resolution structural data of large biomolecules by constructing atomic models with experiment-targeted parallel cascade selection simulations. *Sci. Rep.* **6,** 29360 (2016).
34. Pande, V. S., Beauchamp, K. & Bowman, G. R. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **52,** 99–105 (2010).
35. Chodera, J. D. & Noé, F. Markov state models of biomolecular conformational dynamics. *Current Opinion in Structural Biology* **25,** 135–144 (2014).
36. Pande, V. S., Beauchamp, K. & Bowman, G. R. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **52,** 99–105 (2010).
37. Kitao, A., Hirata, F. & Gō, N. The effects of solvent on the conformation and the collective motions of protein: Normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. *Chem. Phys.* **158,** 447–472 (1991).
38. Molgedey, L. & Schuster, H. G. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.* **72,** 3634–3637 (1994).
39. Pérez-Hernández, G., Paul, F., Giorgino, T., De Fabritiis, G. & Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **139,** 7653 (2013).
40. Naritomi, Y. & Fuchigami, S. Slow dynamics in protein fluctuations revealed by time-

structure based independent component analysis: The case of domain motions. *J. Chem. Phys.* **134,** (2011).

41. Bowman, G. R., Beauchamp, K. A., Boxer, G. & Pande, V. S. Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.* **131,** 124101 (2009).
42. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28,** 129–137 (1982).
43. Prinz, J. H. *et al.* Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.* **134,** (2011).
44. Doerr, S. & de Fabritiis, G. On-the-Fly Learning and Sampling of Ligand Binding by High-Throughput Molecular Simulations. *J. Chem. Theory Comput.* **10,** 2064–2069 (2014).
45. Beauchamp, K. A. *et al.* MSMBuilder2: Modeling conformational dynamics on the picosecond to millisecond scale. *J. Chem. Theory Comput.* **7,** 3412–3419 (2011).
46. Izrailev, S. *et al.* Steered Molecular Dynamics. *Comput. Mol. Dyn. Challenges, Methods, Ideas SE - 2* **4,** 39–65 (1999).
47. Harada, R. & Kitao, A. Nontargeted parallel cascade selection molecular dynamics for enhancing the conformational sampling of proteins. *J. Chem. Theory Comput.* **11,** 5493–5502 (2015).
48. Carrillo, W., Garcia-Ruiz, A., Recio, I. & Moreno-Arribas, M. V. Antibacterial activity of hen egg white lysozyme modified by heat and enzymatic treatments against oenological lactic acid bacteria and acetic acid bacteria. *J Food Prot* **77,** 1732–1739 (2014).
49. Ming, D. & Wall, M. E. Quantifying allosteric effects in proteins. *Proteins Struct. Funct. Genet.* **59,** 697–707 (2005).
50. Cheetham, J. C., Artymiuk, P. J. & Phillips, D. C. Refinement of an enzyme complex with inhibitor bound at partial occupancy: Hen egg-white lysozyme and tri-N-acetylchitotriose at 1.75 Å resolution. *J. Mol. Biol.* **224,** 613–628 (1992).
51. Ishikawa, T. *et al.* A theoretical study of the two binding modes between lysozyme and tri-NAG with an explicit solvent model based on the fragment molecular orbital method. *Phys. Chem. Chem. Phys.* **15,** 3646 (2013).
52. Takemura, K. *et al.* Free-energy analysis of lysozyme–triNAG binding modes with all-atom molecular dynamics simulation combined with the solution theory in the energy representation. *Chem. Phys. Lett.* **559,** 94–98 (2013).

53. Zhong, Y. & Patel, S. Binding structures of tri- *N* -acetyl-β-glucosamine in hen egg white lysozyme using molecular dynamics with a polarizable force field. *J. Comput. Chem.* **34,** 163–174 (2013).
54. Kirschner, K. N. *et al.* GLYCAM06: A generalizable biomolecular force field. carbohydrates. *J. Comput. Chem.* **29,** 622–655 (2008).
55. Pronk, S. *et al.* GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29,** 845–854 (2013).
56. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18,** 1463–1472 (1997).
57. Van Gunsteren, W. F. & Berendsen, H. J. C. A Leap-frog Algorithm for Stochastic Dynamics. *Mol. Simul.* **1,** 173–185 (1988).
58. Swope, W. C., Andersen, H. C., Berens, P. H. & Wilson, K. R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* **76,** 637–649 (1982).
59. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126,** 14101 (2007).
60. Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **81,** 511–519 (1984).
61. Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* **31,** 1695–1697 (1985).
62. 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,** 3684–3690 (1984).
63. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52,** 7182–7190 (1981).
64. Martyna, G. J., Tuckerman, M. E., Tobias, D. J. & Klein, M. L. Explicit reversible integrators for extended systems dynamics. *Mol. Phys.* **87,** 1117–1157 (1996).
65. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **14,** 33–38 (1996).
66. Laskowski, R. A. & Swindells, M. B. LigPlot+: Multiple Ligand–Protein Interaction Diagrams for Drug Discovery. *J. Chem. Inf. Model.* **51,** 2778–2786 (2011).
67. Matsumura, I. & Kirsch, J. F. Is aspartate 52 essential for catalysis by chicken egg white lysozyme? The role of natural substrate-assisted hydrolysis. *Biochemistry* **35,**

1881–1889 (1996).

68. Noguchi, S., Miyawaki, K. & Satow, Y. Succinimide and isoaspartate residues in the crystal structures of hen egg-white lysozyme complexed with tri-N-acetylchitotriose. *J. Mol. Biol.* **278,** 231–238 (1998).
69. Ohmura, T., Ueda, T., Ootsuka, K., Saito, M. & Imoto, T. Stabilization of hen egg white lysozyme by a cavity-filling mutation. *Protein Sci.* **10,** 313–320 (2001).
70. Kamiya, N., Yonezawa, Y., Nakamura, H. & Higo, J. Protein-inhibitor flexible docking by a multicanonical sampling: Native complex structure with the lowest free energy and a free-energy barrier distinguishing the native complex from the others. *Proteins Struct. Funct. Genet.* **70,** 41–53 (2008).
71. Patriksson, A. & van der Spoel, D. A temperature predictor for parallel tempering simulations. *Phys. Chem. Chem. Phys.* **10,** 2073 (2008).
72. Saxton, M. J. Single-particle tracking: the distribution of diffusion coefficients. *Biophys. J.* **72,** 1744–53 (1997).
73. Horton, M., Charras, G. & Lehenkari, P. ANALYSIS OF LIGAND–RECEPTOR INTERACTIONS IN CELLS BY ATOMIC FORCE MICROSCOPY. *J. Recept. Signal Transduct.* **22,** 169–190 (2002).
74. Suárez, E., Adelman, J. L. & Zuckerman, D. M. Accurate Estimation of Protein Folding and Unfolding Times: Beyond Markov State Models. *J. Chem. Theory Comput.* **12,** 3473–3481 (2016).
75. Zhang, B. W. *et al.* Simulating Replica Exchange: Markov State Models, Proposal Schemes, and the Infinite Swapping Limit. *J. Phys. Chem. B* **120,** 8289–8301 (2016).
76. Ogata, M. *et al.* A novel transition-state analogue for lysozyme, 4-O-??-Tri-N-acetylchitotriosyl moranoline, provided evidence supporting the covalent glycosyl-enzyme intermediate. *J. Biol. Chem.* **288,** 6072–6082 (2013).
77. Swanson, J. M. J., Henchman, R. H. & McCammon, J. A. Revisiting Free Energy Calculations: A Theoretical Connection to MM/PBSA and Direct Calculation of the Association Free Energy. *Biophys. J.* **86,** 67–74 (2004).
78. Gohlke, H. & Case, D. A. Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J. Comput. Chem.* **25,** 238–250 (2004).
79. Homeyer, N. & Gohlke, H. Free Energy Calculations by the Molecular Mechanics Poisson−Boltzmann Surface Area Method. *Mol. Inform.* **31,** 114–122 (2012).
80. Lee, M. S. & Olson, M. A. Calculation of Absolute Protein-Ligand Binding Affinity Using Path and Endpoint Approaches. *Biophys. J.* **90,** 864–877 (2006).

81. Gilson, M. K., Rashin, A., Fine, R. & Honig, B. On the calculation of electrostatic interactions in proteins. *J. Mol. Biol.* **184,** 503–516 (1985).
82. Koehl, P., Poitevin, F., Orland, H. & Delarue, M. Modified Poisson–Boltzmann equations for characterizing biomolecular solvation. *J. Theor. Comput. Chem.* **13,** 1440001 (2014).
83. Spiliotopoulos, D. *et al.* dMM-PBSA: A New HADDOCK Scoring Function for Protein-Peptide Docking. *Front. Mol. Biosci.* **3,** 46 (2016).
84. Beveridge, D. L. & DiCapua, F. M. Free Energy Via Molecular Simulation: Applications to Chemical and Biomolecular Systems. *Annu. Rev. Biophys. Biophys. Chem.* **18,** 431–492 (1989).
85. Straatsma, T. P. & McCammon, J. A. Multiconfiguration thermodynamic integration. *J. Chem. Phys.* **95,** 1175 (1991).
86. Woo, H.-J. & Roux, B. Calculation of absolute protein-ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 6825–30 (2005).
87. Kollman, P. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* **93,** 2395–2417 (1993).
88. Simonson, T., Archontis, G. & Karplus, M. Continuum Treatment of Long-Range Interactions in Free Energy Calculations. Application to Protein−Ligand Binding. *J. Phys. Chem. B* **101,** 8349–8362 (1997).
89. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23,** 187–199 (1977).
90. 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).
91. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314,** 141–151 (1999).
92. Harada, R., Nakamura, T. & Shigeta, Y. Sparsity-weighted outlier FLOODing (OFLOOD) method: Efficient rare event sampling method using sparsity of distribution. *J. Comput. Chem.* **37,** 724–738 (2016).
93. Kussie, P. H. *et al.* Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science (80-.).* **274,** 948–953 (1996).
94. Zwier, M. C. *et al.* Efficient Atomistic Simulation of Pathways and Calculation of Rate Constants for a Protein-Peptide Binding Process: Application to the MDM2 Protein and an Intrinsically Disordered p53 Peptide. *J. Phys. Chem. Lett.* **7,** 3440–3445 (2016).

95. Chi, S. W. *et al.* Structural details on mdm2-p53 interaction. *J. Biol. Chem.* **280,** 38795–38802 (2005).
96. Joseph, T. L., Lane, D. & Verma, C. Stapled peptides in the p53 pathway: Computer simulations reveal novel interactions of the staples with the target protein. *Cell Cycle* **9,** 4560–4568 (2010).
97. Moreira, I. S., Fernandes, P. A. & Ramos, M. J. Protein-protein recognition: A computational mutagenesis study of the MDM2-P53 complex. *Theor. Chem. Acc.* **120,** 533–542 (2008).
98. Zondlo, S. C., Lee, A. E. & Zondlo, N. J. Determinants of specificity of MDM2 for the activation domains of p53 and p65: Proline27 disrupts the MDM2-binding motif of p53. *Biochemistry* **45,** 11945–11957 (2006).
99. Schon, O., Friedler, A., Bycroft, M., Freund, S. M. . & Fersht, A. R. Molecular Mechanism of the Interaction between MDM2 and p53. *J. Mol. Biol.* **323,** 491–501 (2002).
100. Dastidar, S. G., Lane, D. P. & Verma, C. S. Multiple peptide conformations give rise to similar binding affinities: Molecular simulations of p53-MDM2. *J. Am. Chem. Soc.* **130,** 13514–13515 (2008).
101. Morris, G. M. & Lim-Wilby, M. Molecular docking. *Methods Mol. Biol.* **443,** 365–382 (2008).
102. Pagadala, N. S., Syed, K. & Tuszynski, J. Software for molecular docking: a review. *Biophys. Rev.* **9,** 91–102 (2017).
103. Mezei, M. A new method for mapping macromolecular topography. *J. Mol. Graph. Model.* **21,** 463–472 (2003).
104. Janin, J. Welcome to CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins: Structure, Function and Genetics* **47,** 257 (2002).
105. Ritchie, D. W. & Kemp, G. J. L. Protein docking using spherical polar Fourier correlations. *Proteins Struct. Funct. Genet.* **39,** 178–194 (2000).
106. Pierce, B. G. *et al.* ZDOCK server: Interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics* **30,** 1771–1773 (2014).
107. Wu, G., Robertson, D. H., Brooks, C. L. & Vieth, M. Detailed analysis of grid-based molecular docking: A case study of CDOCKER?A CHARMm-based MD docking algorithm. *J. Comput. Chem.* **24,** 1549–1562 (2003).
108. Ding, F., Yin, S. & Dokholyan, N. V. Rapid Flexible Docking Using a Stochastic Rotamer Library of Ligands. *J. Chem. Inf. Model.* **50,** 1623–1632 (2010).

109. Ding, F. & Dokholyan, N. V. Incorporating backbone flexibility in MedusaDock improves ligand-binding pose prediction in the CSAR2011 docking benchmark. *J. Chem. Inf. Model.* **53,** 1871–1879 (2013).
110. Chen, H. F. & Luo, R. Binding induced folding in p53-MDM2 complex. *J. Am. Chem. Soc.* **129,** 2930–2937 (2007).

Acknowledgements

First and foremost, I would like to express my sincere gratitute to my supervisor, Prof. Akio Kitao, for non-stop support of my PhD study and research, for all of his encouragement, inspiration and enthusiasm. I appreciate his patience to spend his valuable time, even in his busiest days, to teach me to how to do good in research, to guide me how to solve the problems in research.

Many thanks to Dr. Kazuhiro Takemura for assisting me in all technical problems in research and helping me in daily life, especially Japanese language. Many thanks to all of Kitao's lab members and alumni: Dr. Yu Yamamori, Dr. Hisham Dokainish, Dr. Jacob Swadling, Dr. Akihiro Hata, Dr. Ai Shinobu, Chika Sato San for valuable comments, suggestions and advices in my works. My best regards to Kitao's lab secretaries Ms. Iwasa and Ms. Shibayama to make everything ready, especially paperworks.

I greatly appreciate the Hitachi Scholarship Program of the Hitachi Global Foundation for the financial support to make my PhD dream come true, and to help my curiosity about Japan and Japanese culture, more clearer. I would like to show my respect and thankful to Kazuyuki Miyanaga San, Masaaki Kawamoto San and Tamami Ono San for all of their helps, advices during my stay in Japan.

I would like to send my best acknowledgments to the System B supercomputer in Institute of Solid State Physics of The University of Tokyo, supercomputers at RCCS, and local cluster Artemis in Kitao's Lab. Without them, I cannot finish this dissertation.

I also would like to say a heartfelt "thank you" to my Mother, my deceased Father, my sweetest older Sister, who always believe and encourage me to follow my dream in science.

And finally last but not least to all Vietnamese students and alumni of The University of Tokyo in Kashiwa and Hongo campus area who make my life in Japan more exciting when being away from home.