Hydration layer dynamics and association mechanisms of food and antifreeze proteins

A Molecular Dynamics and Transition Path Sampling study

Brotzakis, Z.F.

Publication date
2017

Document Version
Other version

License
Other

Chapter 2

Methods

Molecular simulations are central to the computation of equilibrium and dynamic properties of classical many body systems, as well as in bridging microscopic with macroscopic observables [1–4] that could be of both experimental and basic importance.

The work of Metropolis et al. [5] in 1953 led to the foundation of Monte Carlo simulations (MC) and introduced the era of molecular simulations. A few years later application of the MC algorithm to a Lennard-Jones potential [6] could quantitatively capture thermodynamic properties. The direct comparison of these computed thermodynamic properties to liquid argon experiments manifested the predictive power of molecular simulations. A few years later in 1957 the Molecular Dynamics (MD) technique was introduced by Alder [7] ; a powerful technique that could capture the dynamics of a system by integrating Newton's equation of motion.

Nowadays Molecular Dynamics is routinely [3] used to explore the conformational dynamics and transitions of complex biological systems, e.g. protein folding, and protein association. Capturing the folding process by brute force Molecular Dynamics simulations of small proteins in the order of microseconds is nowadays tractable by using parallel computing or by use of GPUs [8, 9]. The development of the Anton machine enabled reaching long timescales. The Anton machine [10] is a massively parallel MD specific hardware capable of simulating several milliseconds per day. Many biological transitions occur on that timescale, such as drug binding to proteins, small protein folding [11, 12] etc.

However, natural systems undergo rare but important physical or chemical transitions (transformations) between stable states. For example, large protein folding, conformational changes of molecules, diffusion in solids, crystal nucleation are processes that occur at the millisecond to seconds timescales. To obtain the long time

dynamics of these systems one needs to capture a statistically important amount of these rare transitions, which is intractable even for the Anton machine. One way to overcome this problem is to perform rare event simulations that can adequately sample rare transitions. In this thesis we perform Transition Path Sampling to sample rare transitions between stable states, to identify on pathway intermediates, and understand the mechanism of the transition.

The chapter is organized as follows. We first introduce the Molecular Dynamics technique, an important tool of this thesis. Then we briefly present the Transition Path Sampling algorithm.

2.1 Molecular Dynamics

Classical Molecular Dynamics is a method that solves Newton's deterministic equations of motion. It uses the gradient of the potential energy of a system $V(r)$ to evaluate the trajectory of N atoms, given their initial conditions.

$$F(t) = ma(t) \tag{2.1}$$

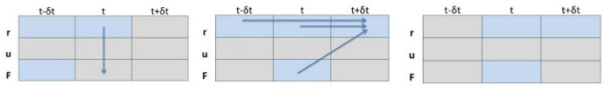
$$u(t) = \frac{dr(t)}{dt} \tag{2.2}$$

$$a(t) = \frac{du(t)}{dt} \tag{2.3}$$

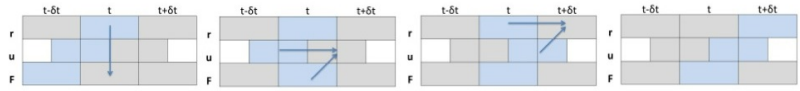
Given the potential energy $V_i(r)$, for each atom i of mass m_i, the force acting on the atom i at time t is given by the gradient of the potential $F_i(r(t)) = -\nabla V_i(r(t))$. Subsequently, one uses an integration algorithm to estimate the positions $r(t+\Delta t)$ and velocities $u(t+\Delta t)$ at time $t+\Delta t$.

There exist many integration algorithms to propagate Newton's equation of motion [13, 14]. Since Newton's equations of motion are time reversible, and preserve the phase space volume, good integrators are considered ones that are time reversible and cause minimal energy drift, therefore, conserving the equilibrium distributions in phase space. Integrators with those properties which are commonly used in Molecular Dynamics simulations are the Verlet, Velocity-Verlet and Leap-Frog algorithms.

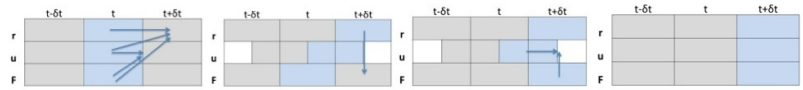
The Verlet algorithm can be derived using a Taylor expansion about $r(t)$. One can sum up the Taylor expansions at $t+\Delta t$ and $t-\Delta t$ shown below to retrieve the equation for advancing the positions in time, as shown in eq 2.6. The advancement of velocities can be retrieved by subtracting eq. 2.4 and eq. 2.5 and is shown in eq. 2.7.



(a)



(b)



(c)

Figure 2.1.1: Three different forms of integrating algorithms. a) Verlet, b) Leap-Frog and c) Velocity-Verlet. Schematic from Allen & Tildesley, Computer Simulation of liquids [1]

$$r(t+\Delta t) = r(t) + \dot{r}\Delta t + \frac{1}{2}\ddot{r}\Delta t^2 + \frac{1}{3}\dddot{r}\Delta t^3 + O(\Delta t^4) \tag{2.4}$$

$$r(t-\Delta t) = r(t) - \dot{r}\Delta t + \frac{1}{2}\ddot{r}\Delta t^2 - \frac{1}{3}\dddot{r}\Delta t^3 + O(\Delta t^4) \tag{2.5}$$

$$r(t+\Delta t) = 2r(t) - r(t-\Delta t) + \frac{F(t)}{m}\Delta t^2 + O(\Delta t^4) \tag{2.6}$$

$$u(t+\Delta t) = \frac{1}{2\Delta t}\left(r(t+\Delta t) - r(t-\Delta t) + O(\Delta t^2)\right) \tag{2.7}$$

As shown in Fig. 2.1.1a, to propagate the positions, the Verlet algorithm uses the positions at time t and $t-\Delta t$ and the forces at t. As shown in eq. 2.6 the Verlet algorithm propagates r about $r(t)$ symmetrically in time thus making it time reversible [1], and simultaneously preserves linear momentum, thus conserving the total energy. However, the algorithm is subject to some computational imprecision since the velocities are calculated with an error of the order Δt^2. Similar to the Verlet algorithm is the Leap-Frog algorithm, shown in equations 2.8 and 2.9.

$$r(t+\Delta t) = r(t) - u(t+\frac{1}{2}\Delta t)\Delta t \tag{2.8}$$

$$u(t+\frac{1}{2}\Delta t) = u(t-\frac{1}{2}\Delta t) + \frac{F(t)}{m}\Delta t \tag{2.9}$$

The positions at time t as well as the velocities at time $t+\frac{\Delta t}{2}$ are used to propagate the positions, whereas the velocities at time $t+\frac{\Delta t}{2}$ are calculated using the forces at time t and the velocities at time $t-\frac{\Delta t}{2}$, as shown in Fig. 2.1.1b.

In the Leap-Frog algorithm, one needs to calculate the velocities at time t in an extra step, using $u(t-\frac{1}{2}\Delta t)$ and $u(t+\frac{1}{2}\Delta t)$. In order to avoid this extra step, the Velocity-Verlet algorithm stores the positions, velocities and forces at the same time step (equations 2.10 and 2.11) while minimizing the round off error.

$$r(t+\Delta t) = r(t) + u(t)\Delta t + \frac{F(t)}{2m}\Delta t^2 \tag{2.10}$$

$$u(t+\Delta t) = u(t) + \frac{F(t+\Delta t)+F(t)}{2m}\Delta t \tag{2.11}$$

All the above algorithms are time reversible and conserve phase space volume in the limit of small time steps. However, since there is no perfect integrating algorithm, there exists a small non-zero error in the calculation of the positions, velocities and forces. A small perturbation in the initial conditions can lead to a deviation from the "true" trajectory that scales exponentially with time according to the Lyapunov exponent λ. Although Molecular Dynamics does not predict the true physical trajectory, even not in principle, it provides statistically trustworthy trajectory predictions [2].

Molecular Dynamics simulations can simulate typically systems of thousands to millions of atoms, using boxes with a linear size of a few tens of nanometres. Such finite size systems cause artifacts on both dynamics and thermodynamics, because of the large fraction of atoms close to the boundaries. If the system had free or wall boundaries, the fraction of particles at the boundary would scale as $N^{-\frac{1}{3}}$, leaving a significant amount of molecules being influenced by the finite size of the system, influencing its dynamic and thermodynamic properties. To simulate bulk properties, periodic boundary conditions (PBC) are used. These periodic boundaries mimic infinite bulk surroundings, where the periodic box is replicated an infinite amount of times into all directions [1, 2]. However, one needs to be careful in the choice of box size, since periodicity rules out fluctuations that are of longer length than the box size. This effect is critical in phase transitions, where long range density fluctuations come into play. In addition, to simulate for instance protein folding transitions or protein association,

one needs to construct a box with a size greater than the cutoffs of the short ranged interactions so that atoms do not interact with their own image, and at the same time large enough to accommodate the transition that one wants to study. For example, for the association of two middle sized proteins ($R_g \approx 2$ nm each) in water, a cubic box of length 10 nm and about 100.000 atoms would be needed in order to capture dissociation events that require to visit a dissociated state with a minimum protein-protein distance of at least 1 nm. Moreover, when using PBCs short ranged atomic interactions are calculated only between nearest periodic neighbours (the minimum image convention). Since the slowest part of the simulation involves the force calculations which scales as N^2, where N is the number of particles in the box, one wants to have the box size to the absolute minimum for more efficient calculations. Hence using truncated octahedron boxes that reduce the solvent molecules about 30 % are commonly used in MD simulations.

2.1.1 Molecular Dynamics of solvated proteins

All atom Force Field models

The propagation of positions and velocities in the Verlet integration algorithm requires as input the force acting on each atom. Since the force on each atom can be obtained by the gradient of the potential energy at each atom, it follows that an atomic description of the potential energy is essential for MD. A force field is a set of classical interatomic semi-empirical potential energy functions that describes the interactions between atoms. Common force fields in biomolecular simulations are the AMBER [15], OPLSAA [16], CHARMM [17], and GROMOS [18]. The basic form of a force field is shown in eqs. 2.12-2.14.

$$V_{tot} = V_{cov} + V_{noncov} = \left[V_{bond} + V_{angle} + V_{dih}\right] + \left[V_{vdw} + V_{el}\right] \tag{2.12}$$

$$\begin{aligned} V_{cov} = & \sum_{i,j} \frac{1}{2} k_b^{ij} \left(r_{ij} - r_{ij}^0\right)^2 + \\ & \sum_{i,j,k} \frac{1}{2} k_a^{ijk} \left(\theta_{ijk} - \theta_{ijk}^0\right)^2 + \\ & \sum_{i,j,k,l} \frac{1}{2} k_d^{ijkl} \left(1 + cos\left(n\phi_{ijkl} - \phi_{ijkl}^0\right)\right) \end{aligned} \tag{2.13}$$

$$V_{noncov} = \sum_{LJ} k_{ij} \left[\left(\frac{C_{ij}^A}{r_{ij}}\right)^{12} - \left(\frac{C_{ij}^B}{r_{ij}}\right)^6\right] + \sum_{el} \frac{q_i q_j}{4\pi\varepsilon_0 r ij} \tag{2.14}$$

The covalent bonds and angle terms are described harmonically, whereas the dihedrals by a periodic function. In eq 2.13, r_{ij} is the bond length between atoms i and j, θ_{ijk} is the valence angle between atoms i, j, k and ϕ_{ijkl} is the dihedral angle between atoms i,j,k,l and n is the multiplicity. The corresponding bond, angle and dihedral equilibrium values are r^0_{ij}, θ^0_{ijk}, ϕ^0_{ijkl} with force constants k_b^{ij}, k_a^{ijk}, k_d^{ijkl} respectively. Non-covalent interactions consist of van der Waals interactions, Pauli repulsions and electrostatic interactions. The 6-12 Lennard-Jones potential parametrizes the Pauli repulsion with a term scaling $\propto r^{-12}$ and the van der Waals attraction with an term $\propto r^{-6}$. The parameters C^A_{ij} and C^B_{ij} are related to the repulsion and attraction between atoms i,j. The electrostatic interactions between atoms i and j are described with a Coulombic term. This term involves the partial atomic charges q_i, q_j, the dielectric permittivity of vacuum ε_0. All of the interactions except for the electrostatic, decay within the box size. However, since the electrostatic interactions scale $\propto r^{-1}$, truncation and tail correction of the electrostatic potential is not a good strategy, as the tail correction energy term will diverge for any potential that decays slower than $\propto r^{-3}$.

A way around this problem is provided by methods that split the sum of electrostatic interactions into a short and a long term. The short range term is solved in real space, and the long range term in reciprocal space. A commonly used method is the Particle Mesh Ewald approach in which the electrostatic interactions are splitted into short and long range terms. Up to a cuttoff, the electrostatic contribution is calculated from the particle-particle interactions whereas for greater distances the charges are put on a grid, and from the grid density the Poisson equation is solved efficiently using Fast Fourier Transform techniques [13]. A similar often used technique is Particle Particle /Particle Mesh (PPPM) [13].

The accuracy of force-fields depends on the target data used to optimize the force-fields parameters. These data include QM calculations as well as experimental measurements. Bond and angle equilibrium constants as well as the multiplicity and phase are obtained from QM calculations of geometries. The vibrational spectra are used to adjust the force constants [19]. Optimization of the non bonded interactions is pivotal for reproducing the thermodynamic properties, as well as the structure of biomolecules. Charge determination methods such as the QM electrostatic potential (ESP) where the charges are optimized to reproduce the QM determined ESP is a standard approach to reproduce well the electronic properties of molecules. Lennard-Jones parameters are adjusted in order to reproduce experimentally determined targeted thermodynamic properties such as the heat of vaporization, density, compressibilities, heat capacities and free energy of solvation [20–22]. Van der Waals interactions can be also obtained using ab initio calculations [23].

In order to make molecular dynamic simulations computationally efficient, one

can constrain the fast vibrational degrees of freedom, thus allowing an increase in the MD time step. In MD simulations these constraints are usually enforced by using the method of Lagrange multipliers. The corresponding algorithms use a Lagrangian including constraints and differ in the way they solve the modified equations of motion to obtain the Lagrange multipliers and therefore the new positions and velocities. Some of the most common constraints algorithms are SHAKE [1], RATTLE [24], LINCS [25], and SETTLE [26]. Another way would be to use the multiple time step approach, by using short time step for the bond vibrations and longer time step for the long range attraction forces (e.g RESPA algorithm [27]).

A very time consuming part of the MD algorithm is the neighbour search of particle i through all possible particles j, in order to determine the particles j which reside within a cutoff distance from i and carry out the force calculation F_{ij}. This procedure, scales as N^2. In order to reduce this computational cost, Verlet [28] suggested the use of neighbour lists. In this scheme, one does not update the neighbours of i at every time step but with some lower frequency. Between the updates of the neighbour list the forces are calculated between particles in the stored neighbour list.

As mentioned earlier, Newton's equation of motion conserve the total energy of the system also known as the Hamiltonian. Thus, MD samples the microcanonical ensemble (NVE). Constant temperature simulations are achieved by using thermostats, such as the v-rescale [29], Nose-Hoover [30] or Andersen [31] thermostats, that allow to sample the canonical ensemble (NVT). Moreover, constant pressure calculations are achieved by using barostats , such as the Parrinello-Rahman [32] barostat, that allow sampling of the isobaric isothermal ensemble (NPT).

For the MD simulations we used Gromacs as an MD engine due to its efficiency and documentation on biomolecular simulations [33].

Water Force Field models

Accurate treatment of the condensed aqueous environment is central in biomolecular simulations. This treatment is performed using explicit and implicit water force-fields. Some famous explicit water force-fields are the TIP3P [34], TIP4P [35], SPC [36] and TIP4P/2005 [37]. These force-fields are parametrized such that they reproduce static experimental properties such as the radial distribution function of water, as well as dynamic ones, such as the diffusion constant. SPC and TIP3P have interaction sites assigned to each of its atoms. Although SPC and TIP3P have slightly different tetrahedral angles (109.4 $^\circ$ and 104.5 $^\circ$ respectively), their charge distribution does not differ much. TIP4P has four interaction sites with an extra charge placed on a dummy atom, which improves the electrostatic distribution at the cost of computational time [38]. The TIP4P/2005 model has almost the same charge distribution and angle

with TIP4P. Their difference is that the TIP4P has been parametrized to target the enthalpy of vaporization. Essentially, TIP4P/2005 estimates better all properties but the ones related to the gas phase, compared to TIP4P [38].

2.2 Rare Event methods

Many high dimensional complex systems in chemistry, physics and biology undergo rare physical or chemical transformations between metastable states. Such rare transitions include protein folding, protein association, ion dissociation, cluster isomerization, crystal nucleation diffusion in solids, etc.

A variety of methods exists that alleviate the exponential time scale problem (e.g umbrella sampling [39], local elevation [40], conformational flooding [41], adaptive bias force [42], metadynamics [43]) and give access to an ergodic sampling of the phase space. However, these methods are based on implementing a bias along an order parameter, which alters the dynamics, gives up kinetic information and at the same time their efficiency depends on the quality of the order parameter as a reaction coordinate. On the other hand, path based methods do not require a reaction coordinate as an input, but only the state definitions and some initial chain of states connecting them. Methods such as the Nudge Elastic Band (NEB) [44] or the zero temperature string method [45] aim to find the minimum energy path (MEP) along one transition. Although these approaches give useful insight into the activation energy and the location of the saddle point, they do not contain dynamical information, since the evolution between states results from a minimization of an objective function. This problem was addressed in another category of path based methods, the so called action-based methods, which use the Hamilton's principle of least action to obtain true dynamical trajectories connecting the two states [46, 47]. Finally, Transition Path Sampling [48, 49] is a trajectory based method, based on an importance sampling strategy on trajectory space, giving insight into both mechanistic and dynamical information. We will discuss this method in more detail in the following section.

2.3 Transition Path Sampling

Transition Path Sampling (TPS) is based on the idea that for a two state system, an infinitely long trajectory crosses the barrier infinite amount of times. The collection of all possible pathways that cross the barrier and connect the stable states is denoted the path ensemble and can be sampled by a MC random walk in the trajectory space. TPS, similar to many other trajectory based methods, needs a good definition of the

reactant and the product state but requires no prior information on the reaction coordinate. Given an initial path connecting the two states and using a Metropolis-Hastings criterion based on detailed balance one can generate a Markov chain of true dynamical pathways that maintain the equilibrium path ensemble distribution. There are many schemes to generate new pathways. In this thesis the one-way flexible shooting scheme has been employed and will be discussed later.

In TPS a path consists of a sequence of microstates at each time slice $\mathbf{x}(T) \equiv \{x_0, x_{\Delta t}, ..., x_{i\Delta t}, ..., x_T\}$, where $x_{i\Delta t} = \{r_{i\Delta t}, p_{i\Delta t}\}$ comprises the coordinates $r_{i\Delta t}$ and momenta $p_{i\Delta t}$ of all the atoms of the system at each time step $i = 0, 1, ..., N$ where $N = \frac{T}{\Delta t} + 1$ (Fig. 2.3.1). One can express the probability $P(\mathbf{x}(T))$ of a pathway of duration T in a statistical mechanical way. $P(\mathbf{x}(T))$ depends on the initial conditions and the underlying dynamics of the system according to eq 2.15.

$$P(\mathbf{x}(T)) = \rho_{x0} \prod_{i=0}^{\frac{T}{\Delta t}-1} p(x_{i\Delta t} \rightarrow x_{i+1\Delta t})/Z(T) \tag{2.15}$$

$p(x_{i\Delta t} \rightarrow x_{i+1\Delta t})$ is the short time probability to evolve from $x_{i\Delta t}$ to $x_{i+1\Delta t}$ and ρ_{x0} is the initial conditions distribution that scales according to the standard Boltzmann factor in the canonical ensemble. The normalization constant or partition function $Z(T)$ is given by eq 2.16, where the path-integral $\int D\mathbf{x}$ denotes summation over all pathways $\mathbf{x}$. For an infinite and discretized path this corresponds to summation over all phase space points.

$$Z(T) \equiv \int P(\mathbf{x}) D\mathbf{x} \tag{2.16}$$

The transition path ensemble is the subset of the reactive trajectories, the ones connecting the stable states A and B. Similarly, the probability of a path $P_{AB}(T)$ is given by equations 2.17-2.18.

$$P_{AB}(\mathbf{x}) \equiv h_A(x_0) h_B(x_T) P(\mathbf{x}) Z_{AB}^{-1} \tag{2.17}$$

$$Z_{AB}(T) \equiv \int h_A(x_0) h_B(x_T) P(\mathbf{x}) D\mathbf{x} \tag{2.18}$$

The characteristic functions $h_A(x)$ and $h_B(x)$ equal unity when the system at a configuration x is in stable state A or B respectively and equal zero elsewhere.

Parametrization of the stable states is based on projecting the high dimensional phase space into order parameter(s) which need to fulfil several criteria. Those are : 1) the state definitions should not overlap, 2) the order parameters should distinguish the states and 3) the states should be attractors and commit the trajectory beyond the barrier region with a high probability.

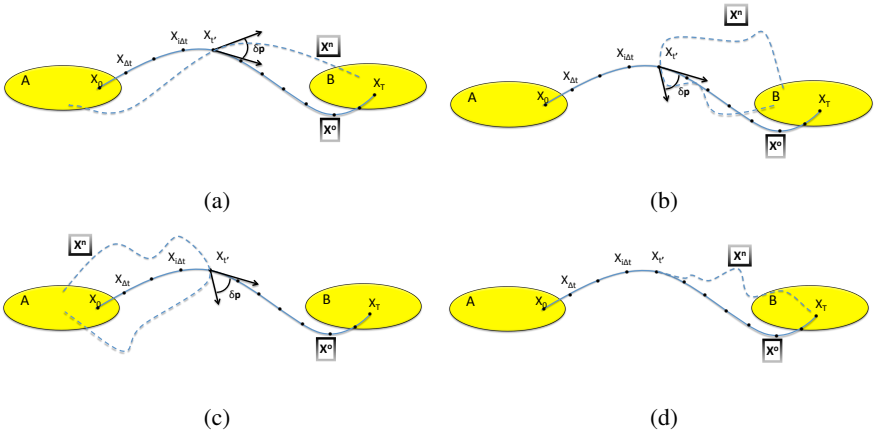


Figure 2.3.1: Schematic representation of the old ($\mathbf{x}^{(o)}$ depicted in solid line) and newly generated paths ($\mathbf{x}^{(n)}$ depicted with dashed lines) a) connecting the two states b,c) not connecting the states and using the two way shooting algorithm by perturbing the momenta at time t'. In d) is described the one-way shooting algorithm where the new path is the old partial path (solid line) plus the newly generated forward partial path (dashed line).

2.3.1 Monte Carlo of trajectories

TPS is a random walk through trajectory space where by construction, trajectories are generated proportionally to their probability in the Transition Path Ensemble (TPE) (eq 2.17). This MC random walk is realized by generating a new trial trajectory and accepting or rejecting it according to the Metropolis criterion. A trial move consists of the generation of a new trajectory $\mathbf{x}^{(n)}$ from an old trajectory $\mathbf{x}^{(o)}$ with a probability $P_{gen}(\mathbf{x}^{(o)} \rightarrow \mathbf{x}^{(n)})$ and the acceptance of the newly generated trajectory with some probability $P_{acc}(\mathbf{x}^{(o)} \rightarrow \mathbf{x}^{(n)})$. Here the superscripts 'n' and 'o' abbreviate the 'new' and 'old' path, respectively. As detailed balance ensures that the probability of the forward move is equal to the probability of the reverse move ($p(o \rightarrow n)$=$p(o \leftarrow n)$), in the TPS framework this translates into eq 2.19

$$P_{AB}(\mathbf{x}^{(o)})P_{gen}(\mathbf{x}^{(o)} \rightarrow \mathbf{x}^{(n)})P_{acc}(\mathbf{x}^{(o)} \rightarrow \mathbf{x}^{(n)}) = P_{AB}(\mathbf{x}^{(n)})P_{gen}(\mathbf{x}^{(n)} \rightarrow \mathbf{x}^{(o)})P_{acc}(\mathbf{x}^{(n)} \rightarrow \mathbf{x}^{(o)}) \tag{2.19}$$

which states that the probability P_{AB} to observe an old path multiplied with the conditional probability of generating and accepting a new path is equal to the probability of observing the old path multiplied with the probability of accepting and generating the old path from the new path. Using eq 2.17, the fact that $h_A(x_0^{(o)}) = h_B(x_T^{(o)}) = 1$ then the acceptance probability can be written as in eq 2.20,

$$\frac{P_{acc}(\mathbf{x}^{(o)} \to \mathbf{x}^{(n)})}{P_{acc}(\mathbf{x}^{(n)} \to \mathbf{x}^{(o)})} = h_A(x_0^{(n)}) h_B(x_T^{(n)}) \frac{P(\mathbf{x}^{(n)}) P_{gen}(\mathbf{x}^{(n)} \to \mathbf{x}^{(o)})}{P(\mathbf{x}^{(o)}) P_{gen}(\mathbf{x}^{(o)} \to \mathbf{x}^{(n)})} \tag{2.20}$$

where P_{gen} is dependent on the specifics of the algorithm that generates the pathways. In order to fulfil the above equation one can apply a Metropolis rule, shown in eq 2.21

$$P_{acc}(\mathbf{x}^{(o)} \to \mathbf{x}^{(n)}) = h_A(x_0^{(n)}) h_B(x_T^{(n)}) \min\left\{1, \frac{P(\mathbf{x}^{(n)}) P_{gen}(\mathbf{x}^{(n)} \to \mathbf{x}^{(o)})}{P(\mathbf{x}^{(o)}) P_{gen}(\mathbf{x}^{(o)} \to \mathbf{x}^{(n)})}\right\} \tag{2.21}$$

where the *min* function returns the smaller of its arguments. This means that if the acceptance probability is one then the move is accepted, whereas if it is lower than one the move is accepted only if the acceptance probability is larger than a random number between 0 and 1.

2.3.2 Shooting move

In the seminal TPS paper [48], the algorithm is based on the shooting move, where old reactive paths are perturbed to lead to new accepted paths. The shooting algorithm works as follows. First one selects a time slice $x_{t'}^{(o)}$ which belongs to the old path $\mathbf{x}^{(o)}(T)$ with some probability $p_{sel}(t', \mathbf{x}^{(o)})$. This time-slice is denoted as the shooting point and in the most straightforward version of the shooting algorithm the shooting point is uniformly selected (it has the same probability along the path). Then one can modify this shooting point by, for example, slightly changing the momenta, followed by a forward and a backward integration of Newton's equations until $t = 0$ and $t = T$. Due to molecular chaos, the new path is expected to diverge exponentially from the old path. If the path connects the two states it can be accepted with a probability according to eq 2.21. The generation probability can be written analytically as the product of probabilities to select a shooting point $p_{sel}(t', \mathbf{x}^{(o)})$, to modify it (e.g. changing the momenta) $p_{gen}(x_{t'}^{(o)} \to x_{t'}^{(n)})$ and the probability to create a partial path [50]. However, since 1) the dynamics conserves the stationary distribution $\rho_{st}(x)$, microscopic reversibility gives $p(x \to y)/p(y \to x) = \rho_{st}(y)/\rho_{st}(x)$ 2) the initial conditions belong to the stationary distribution, $\rho(x) = \rho_{st}(x)$ and 3) the generation and selection probabilities are symmetric, equation 2.21 simplifies to eq 2.22.

$$P_{acc}(\mathbf{x}^{(o)} \to \mathbf{x}^{(n)}) = h_A^{(n)}(0) h_B^{(n)}(T) \min\left\{1, \frac{\rho_{x_{t'}}^{(n)}}{\rho_{x_{t'}}^{(o)}}\right\} \tag{2.22}$$

Complex biomolecular systems often encounter long transition pathways. The length of these pathways is often much longer than the Lyapunov time λ it takes for the molecular chaos to appear. This means that although the dynamics is deterministic (e.g Newtonian dynamics) the system essentially behaves diffusively. In these kind of systems, the two way shooting approach with a perturbation of the momenta will lead to a low acceptance ratio [51] as both the forward and backward paths have to reach the correct stable states, while the new pathway diverges quickly from the previous accepted pathway, and is likely to lose its reactivity. The one way shooting alleviates this problem as it only shoots one way per trial (forward or backward) thus increasing the acceptance ratio as only one partial path has to end in the correct state. The new pathway is accepted if the partial path ends in the correct state B for a forward path and A for a backward path. The new path consists of the old part and the new partial path, see Fig. 2.3.1d. The one-way shooting algorithm uses a stochastic thermostat, exploiting both the natural tendency to decorrelate due to the size of the system as well as the stochasticity introduced to the dynamics by the thermostat. No perturbation of momenta is done at the selected shooting point since this would only disrupt the natural NVE dynamics of the whole trajectory. As in the one-way shooting the new accepted path comprises the old and new partial path, the downside is that it takes more accepted paths before one reaches a path that shares no common points with the previous path, a so-called *decorrelated path*.

For reducing the computational cost of TPS, the flexible path length algorithm was introduced, where transition paths adjust to the length of the transition part, avoiding long dwelling in the stable states [50, 52]. As the selection probability for a shooting point is $p_{sel}(t', \mathbf{x}^{(o)}) = 1/L^{(o)}$, $p_{sel}(t', \mathbf{x}^{(n)}) = 1/L^{(n)}$, the acceptance probability becomes

$$P_{acc}(\mathbf{x}^{(o)} \to \mathbf{x}^{(n)}) = h_A^{(n)}(0) h_B^{(n)}(T) \min\left\{1, \frac{L^{(o)}}{L^{(n)}}\right\} \tag{2.23}$$

where $L^{(o)}$ and $L^{(n)}$ is the length of the old and new transition path respectively. In our implementation of the one-way shooting, the v-rescale [29] stochastic thermostat was used.

Bibliography

[1] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Clarendon Press, New York, NY, USA, 1989).

[2] D. Frenkel and B. Smit, *Understanding Molecular Simulation* (Academic Press, Inc., Orlando, FL, USA, 2001), 2nd ed.

[3] M. Karplus and R. Lavery, Isr. J. Chem. **54**, 1042 (2014).

[4] D. N. Theodorou, Ind. Eng. Chem. Res. **49**, 3047 (2010).

[5] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

[6] W. W. Wood and F. R. Parker, J. Chem. Phys. **27**, 720 (1957).

[7] B. J. Alder and T. E. Wainwright, J. Chem. Phys. **27**, 1208 (1957).

[8] Y. M. Rhee, E. J. Sorin, G. Jayachandran, E. Lindahl, and V. S. Pande, Proc. Natl. Acad. Sci. U.S.A. **101**, 6456 (2004).

[9] T. J. Lane, D. Shukla, K. A. Beauchamp, and V. S. Pande, Curr. Opin. Struc. Biol. **23**, 58 (2013).

[10] D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, et al., Commun. ACM **51**, 91 (2008).

[11] R. O. Dror, A. C. Pan, D. H. Arlow, and D. E. Shaw, *Chapter 20 Probing the Conformational Dynamics of GPCRs with Molecular Dynamics Simulation*, 8 (The Royal Society of Chemistry, 2011).

[12] Y. Shan, E. T. Kim, M. P. Eastwood, R. O. Dror, M. A. Seeliger, and D. E. Shaw, J. Am. Chem. Soc. **133**, 9181 (2011).

[13] R. W. Hockney and E. J. W, *Computer simulation using particles* (McGraw-Hill International Book Co,New York,USA, 1981).

[14] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, J. Chem. Phys. **76**, 637 (1982).

[15] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, J. Am. Chem. Soc. **117**, 5179 (1995).

[16] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-rives, J. Am. Chem. Soc. **118**, 11225 (1996).

[17] A. D. Mackerell, D. Bashford, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, L. Kuchnir, et al., J. Phys. Chem. B **102**, 3586 (1998).

[18] W. R. P. Scott, P. H. Hu, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Kru, and W. F. V. Gunsteren, J. Phys. Chem. A **103**, 3596 (1999).

[19] A. D. Mackerell, J. Comp. Chem. **25**, 1584 (2004).

[20] W. L. Jorgensen, J. Phys. Chem. **90**, 6379 (1986).

[21] A. Warshel, J. Chem. Phys. **53**, 582 (1970).

[22] A. D. Mackerell, J. Wirkiewicz-kuczera, and M. Karplus, J. Am. Chem. Soc. **117**, 11946 (1995).

[23] D. Yin and A. D. Mackerell, J. Phys. Chem. **100**, 2588 (1996).

[24] C. Andersen, J. Comput. Phys. **52**, 24 (1983).

[25] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, J. Comp. Chem. **18**, 1463 (1997).

[26] S. Miyamoto and P. A. Kollman, J. Comp. Chem. **13**, 952 (1992).

[27] D. D. Humphreys, R. A. Friesner, and B. J. Berne, J. Phys. Chem. **98**, 6885 (1994).

[28] L. Verlet, Phys. Rev. **159**, 98 (1967).

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

[30] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).

[31] H. C. Andersen, J. Chem. Phys. **72**, 2384 (1980).

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

[33] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, SoftwareX **1–2**, 19 (2015).

[34] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).

[35] W. L. Jorgensen, J. F. Blake, and J. K. Buckner, Chem. Phys. **129**, 193 (1989).

[36] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans, *Interaction Models for Water in Relation to Protein Hydration* (Springer, Dordrecht, Netherlands, 1981).

[37] J. L. F. Abascal and C. Vega, J. Chem. Phys. **123**, 234505 (2005).

[38] C. Vega and J. L. F. Abascal, Phys. Chem. Chem. Phys **13**, 19663 (2011).

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

[40] P. Chemistry, J. Comput. Aided Mol. Des **8**, 695 (1994).

[41] H. Grubmuller, Phys. Rev. E **52**, 2893 (1995).

[42] E. Darve, D. Rodríguez-Gómez, and A. Pohorille, J. Chem. Phys. **128**, 144120 (2008).

[43] A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. USA **99**, 12562 (2002).

[44] G. Henkelman, B. P. Uberuaga, and H. Jonson, J. Chem. Phys. **113**, 9901 (2000).

[45] W. E, W. Ren, and E. Vanden-Eijnden, Phys. Rev. B. **66**, 052301 (2002).

[46] D. Passerone and M. Parrinello, Phys. Rev. Lett. **87**, 1 (2001).

[47] R. Olender and R. Elber, J. Chem. Phys. **105**, 9299 (1996).

[48] C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, J. Chem. Phys. **108**, 1964 (1998).

[49] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).

[50] P. G. Bolhuis and C. Dellago, Rev. Comp. Chem. **27**, 1 (2010).

[51] P. Bolhuis, J. Phys.: Condens. Matter 15 **15**, 113 (2003).

[52] T. S. van Erp, D. Moroni, and P. G. Bolhuis, J. Chem. Phys. **118**, 7762 (2003).