CHAPTER 2

Methodology

2.1 Computer Simulation

Computer simulation is an useful tool for studying structural, dynamical and thermodynamical properties of various molecular systems and for unravelling the complexity of physical phenomena at molecular-level using the fundamental principles of statistical mechanics, classical and quantum mechanics. By tracking the motions of individual atoms in the system, computer simulations enable us to derive experimentally measurable macroscopic properties of the system of interest (including the equation of state, transport coefficients, and structural and dynamical order parameters) and to probe the precise pathways and energetics of many physical processes. The growing advances in high-performance computing hardwares and improved algorithms continue to push the length- and time-scale boundaries there by enabling us to model larger systems and larger processes. Computer simulation techniques include quantum mechanics-based methods such as ab initio quantum chemical, density functional theory (DFT) and classical mechanics-based methods such as Monte Carlo (MC) and molecular dynamics (MD). In the present thesis, MD and enhanced sampling MD methods are employed to study the conformational dynamics and thermodynamics of fast side chain motions in proteins.

2.2 Molecular Dynamics Simulations

In molecular dynamics (MD) simulations, the atomic forces derived from the empirical potential energy function are used to solve the equations of motion, thereby tracking the dynamical trajectories of individual atoms in the system. Using the principles of statistical mechanics, the macroscopic properties of the system are estimated from the dynamical trajectories. The first MD simulation was performed by Alder and Wainwright in 1957 to study phase transitions in hard sphere systems.[67] In 1964, Aneesur Rahman introduced the concept of realistic potentials into MD simulations to probe the structure, dynamics and energetics of liquid argon.[68] Karplus et al. performed MD simulation of protein bovine pancreatic trypsin inhibitor (BPTI) in vacuum for 9.2ps in 1975. Marked as the beginning of modern computational biochemistry and biophysics, the work of Karplus et al. is being followed by a large number of theoretical investigations of many complex biomolecular systems till date.

2.2.1 Force Field

In MD simulation, the covalent and noncovalent interactions between constituent atoms of a system of interest are described by an empirical potential energy function (also known as force field), $V(\mathbf{r})$, which is given by

$$\begin{split} V(\mathbf{r}) = \frac{1}{2}\sum_{bonds} K_b(b-b_0)^2 + \frac{1}{2}\sum_{angles} K_\theta(\theta-\theta_0)^2 + \frac{1}{2}\sum_{torsions} K_\phi(1+cos(n\phi-\delta)) \\ +\sum_{ij} 4\varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) + \sum_{ij}\frac{q_i q_j}{4\pi\varepsilon r_{ij}} \end{split} \tag{2.1}$$

The first and second terms correspond to bond-stretching and angle-bending, respectively and are treated harmonically, which effectively keeps the bonds and angles near their equilibrium values. The parameters b_0 and θ_0 denote the equilibrium bond length and equilibrium angle, respectively. K_b and K_θ are the force constants associated with the bond and angle terms, respectively. The third term is for the dihedrals, where K_ϕ is the dihedral force constant, n is the multiplicity of the function, ϕ is the dihedral angle and δ is the phase shift.

Nonbonded interactions between pairs of atoms (i,j) are represented by the last term consisting of van der Waals and electrostatic interactions. van der Waals interaction is described by a standard 12-6 Lennard-Jones (LJ) potential (see Figure 2.1). r_{ij} is the distance between the interaction pair. The term r_{ij}^{-12} dominates at short distance contributing to the repulsion between atoms whereas r_{ij}^{-6} dominating at large distance contributes to the attraction between them. The attractive contribution is due to instantaneous dipoles which arise during fluctuations in the electron clouds and repulsive contribution is due to nuclear repulsion and short

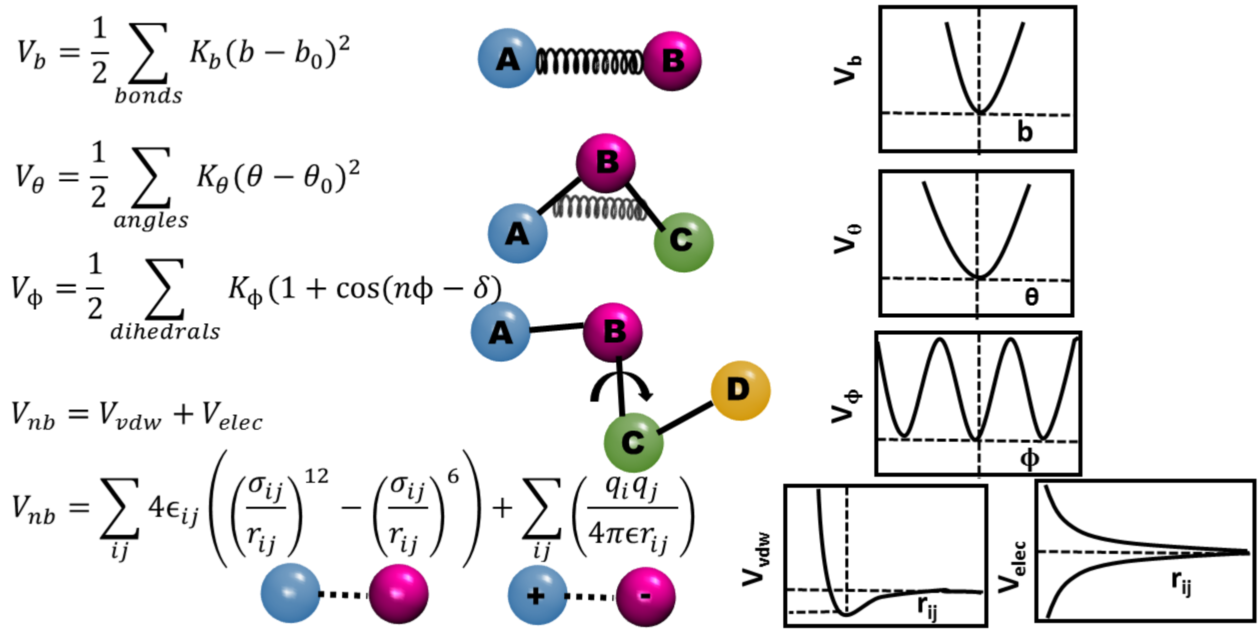
$V_b = \frac{1}{2} \sum_{bonds} K_b(b - b_0)^2$
A
B
V_b
b
$V_\theta = \frac{1}{2} \sum_{angles} K_\theta(\theta - \theta_0)^2$
B
A
C
V_θ
θ
$V_\phi = \frac{1}{2} \sum_{dihedrals} K_\phi(1 + \cos(n\phi - \delta))$
A
B
C
D
V_ϕ
ϕ
$V_{nb} = V_{vdw} + V_{elec}$
$V_{nb} = \sum_{ij} 4\epsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) + \sum_{ij}\left(\frac{q_i q_j}{4\pi\epsilon r_{ij}}\right)$
+
-
V_{vdw}
r_{ij}
V_{elec}
r_{ij}

Figure 2.1: Schematic illustrating the forcefield.

range repulsive forces (overlap forces). ε_{ij} relates to the LJ well depth, σ_{ij} is the distance between i and j at which the minimum LJ interaction energy occurs and is related to van der Waals radius of an atom. The Lorentz-Berthelodt combination rules are used to obtain the necessary LJ parameters for each interaction pair; ε_{ij} values are the geometric mean of ε_{ii} and ε_{jj} values determined for the individual atom types while σ_{ij} values are the arithmetic mean of the σ_{ii} and σ_{jj} values for individual atom types. The electrostatic interaction between the partial atomic charges q$_i$ and q$_j$ with dielectric constant ε is obtained using the coulombic potential.[69–71]

2.2.2 Theory

The microscopic state of a system is defined by the positions and momenta of the atoms constituting the system. The Hamiltonian H of a system of N atoms, which is a sum of kinetic and potential energy functions, is expressed in terms of a set of coordinates $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N)$

and momenta $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_N)$ of atoms as follows

$$H(\mathbf{r},\mathbf{p}) = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + V(\mathbf{r}) \tag{2.2}$$

where m_i is the mass of atom i. The potential energy V($\mathbf{r}$) defines the interatomic interactions in the system and is given in Equation 3.1. The force on each atom is calculated from V using $\mathbf{F} = -\nabla_i V(\mathbf{r})$. The equations of motion that govern the time-evolution of the system are given by the Hamilton's equations of motions.

$$\dot{\mathbf{p}}_i = -\frac{\partial H}{\partial \mathbf{r}_i} = -\frac{\partial V}{\partial \mathbf{r}_i} = \mathbf{F}_i \tag{2.3}$$

$$\dot{\mathbf{r}}_i = \frac{\partial H}{\partial \mathbf{p}_i} = \frac{\mathbf{p}_i}{m_i} \tag{2.4}$$

where $\mathbf{r}_i$, $\mathbf{p}_i$, $\dot{\mathbf{r}}_i$ and $\dot{\mathbf{p}}_i$ represent the position, momentum and their time-derivatives, respectively of the ith particle.

The equations of motion are integrated numerically to obtain the dynamics of the system. If $\mathbf{r}_i(t)$ and $\mathbf{p}_i(t)$ represents the position and momentum of ith particle at time t, then its position and momentum after a time-step   are given by

$$\mathbf{r}_i(t+\ \ ) = \mathbf{r}_i(t) + \mathbf{v}_i(t)\ \ + \frac{1}{2m_i}\mathbf{F}_i(t)\ \ ^2 \tag{2.5}$$

$$\mathbf{p}_i(t+\ \ ) = \mathbf{p}_i(t) + \frac{1}{2m_i}[\mathbf{F}_i(t) + \mathbf{F}_i(t+\ \ )] \tag{2.6}$$

By repeating this procedure for a required number of time-steps, the trajectory of the system of interest can be obtained.[69–72]

2.3 Integration Methods

Given the initial coordinates (structure) and velocities of the particles constituting the system, the dynamics of the system can be obtained by solving Newton's equation of motion. The initial structure can be obtained from experimental techniques (X-ray crystallography, NMR spectroscopy) or by computational modelling. The initial velocities for all the particles are assigned using Maxwell-Boltzmann distribution. It is a Gaussian distribution which provides the probability of a particle i of mass m_i to possess the velocity $\mathbf{v_i}$ at a temperature T

$$p(\mathbf{v}_i) = \sqrt{\frac{m_i}{2\pi k_B T)}} exp\left[-\frac{m_i \mathbf{v_i^2}}{2k_B T}\right]. \tag{2.7}$$

Using these initial conditions, different integration alogrithms including verlet, leap-frog and velocity verlet are used to integrate the equations of motion numerically.[69–71]

2.3.1 Verlet Algorithm

Using verlet algorithm, the new positions are calculated using positions at time t and t-δt and accelerations at time t. The positions at time t+δt and $t - \delta t$ can be written as

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2}(\delta t)^2 \mathbf{a}(t) + ... \tag{2.8}$$

$$\mathbf{r}(t-\delta t) = \mathbf{r}(t) - \delta t \mathbf{v}(t) + \frac{1}{2}(\delta t)^2 \mathbf{a}(t) + .. \tag{2.9}$$

Addition of these two equations gives

$$\mathbf{r}(t+\delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\delta t) + (\delta t)^2 \mathbf{a}(t) \tag{2.10}$$

The velocities are calculated using the following

$$\mathbf{v}(t) = \frac{[\mathbf{r}(t+\delta t) - \mathbf{r}(t-\delta t)]}{2\delta t} \tag{2.11}$$

It is clear from the above equation that velocity can only be computed once $r(t+\delta t)$ is known

Alternatively, the velocities can be estimated at the half-step, $t+\frac{1}{2}\delta t$

$$\mathbf{v}(t+\frac{1}{2}\delta t) = \frac{[\mathbf{r}(t+\delta t) - \mathbf{r}(t)]}{\delta t} \tag{2.12}$$

This is not a self-starting algorithm as the new positions are obtained from the current positions $\mathbf{r}(t)$ and the positions from the previous step $\mathbf{r}(t-\delta t)$.[69–71]

2.3.2 Leap-Frog Algorithm

The leap-frog algorithm uses the following relations

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t+\frac{1}{2}\delta t) \tag{2.13}$$

$$\mathbf{v}(t+\frac{1}{2}\delta t) = \mathbf{v}(t-\frac{1}{2}\delta t) + \delta t \mathbf{a}(t) \tag{2.14}$$

The velocity equation is implemented first and the velocities leap over the coordinates to give the next mid-step values $\mathbf{v}(\mathbf{t}+\frac{1}{2}\delta\mathbf{t})$

The new positions $\mathbf{r}(t+\delta t)$ are dependent on $\mathbf{v}(t+\frac{1}{2}\delta t)$ which in turn are dependent on $\mathbf{v}(t-\frac{1}{2}\delta t)$ and $\mathbf{a}(t)$. The velocities at time t are calculated using

$$\mathbf{v}(t)=\frac{1}{2}[\mathbf{v}(t+\frac{1}{2}\delta t)+\mathbf{v}(t-\frac{1}{2}\delta t)] \tag{2.15}$$

This alogrithm evalutes the velocities at half-integer time steps and uses the velocities to compute the new positions.[69–71]

2.3.3 Velocity-Verlet Algorithm

The velocity verlet method uses the following

$$\mathbf{r}(t+\delta t)=\mathbf{r}(t)+\delta t\mathbf{v}(t)+\frac{1}{2}(\delta t)^2\mathbf{a}(t) \tag{2.16}$$

$$\mathbf{v}(t+\delta t)=\mathbf{v}(t)+\frac{1}{2}\delta t[\mathbf{a}(t)+\mathbf{a}(t+\delta t)] \tag{2.17}$$

Using the velocities and accelerations at time,t the positions at $t+\delta t$ are calculated. The velocities at time $t+\frac{1}{2}\delta t$ are then determined using

$$\mathbf{v}(t+\frac{1}{2}\delta t)=\mathbf{v}(t)+\frac{1}{2}\delta t\mathbf{a}(t) \tag{2.18}$$

New forces are next computed from the current positions, thus giving $a(t+\delta t)$. In the final step, the velocities at $t+\delta t$ are determined using

$$\mathbf{v}(t+\delta t)=\mathbf{v}(t+\frac{1}{2}\delta t)+\frac{1}{2}\delta t\mathbf{a}(t+\delta t) \tag{2.19}$$

2.4 Temperature and Pressure Control

Most of the experimental measurements are made under conditions of constant temperature and/or pressure. The thermodynamic enemble corresponding to such conditions are canonical or NVT ensemble, the isothermal-isobaric or NPT ensemble. MD simulations thus performed using NPT/NVT ensemble can be directly be compared with experimental data.

The temperature of the system is related to the time average of kinetic energy as follows

$$<k>=\frac{3}{2}Nk_BT \tag{2.20}$$

The temperature of the system thus can be altered by scaling the velocities at each step by a factor required to attain desired temperature.

Pressure of the system can be calculated via the virial theorem of Clausius as follows

$$P = \frac{1}{V}\left[Nk_BT - \frac{1}{3}\sum_{i=1}^{N}\sum_{j=i+1}^{N} r_{ij}f_{ij}\right] \tag{2.21}$$

where N is number of atoms, T is the temperature and $f_{ij} = \frac{dv(r_{ij})}{dr_{ij}}$ is the force acting between atoms i and j of the system.[69–71]

2.4.1 Berendsen Thermostat and Barostat

In order to achieve temperature or pressure control, the system is to be coupled with external bath. Berendsen et al. proposed the modification to the equation of motion for the velocities of the atoms, v, to accomplish the coupling

$$\dot{v} = M^{-1}F + \frac{1}{2\tau_T}(\frac{T_B}{T} - 1)v \tag{2.22}$$

Where T_B and T are the reference (temperature of external thermal bath) and instantaneous temperature, respectively and τ_T is the coupling constant (with units of time). The additional term to the equation of motion acts like a frictional force. If the system temperature increases relative to the desired temperature, the frictional force becomes negative damping the motions of the atoms and kinetic energy thus reducing the temperature. If the temperature decreases, the frictional force becomes positive and the energy is then supplied to the system. The coupling constant, τ_T determines the strength of the coupling to the external bath. Larger τ_T indicates weak coupling where the system is steered slowly to the bath temperature and vice-versa. The pressure control equations of motions proposed by Berendsen et al. are

$$\dot{R} = V - \frac{\beta}{3\tau_p}(P_B - P)R \tag{2.23}$$

$$\dot{V} = -\frac{\beta}{\tau_p}(P_B - P)V \tag{2.24}$$

where P_B and P are reference and instantaneous pressure, respectively and beta and τ_P are the isothermal compressibility of the system (with inverse pressure units) and pressure coupling constant (time units), V is the volume of the system. In practice, for temperature control the velocities of the particles are scaled with a factor

$$\lambda = \sqrt{1 + \frac{\delta t}{\tau_T}\left(\frac{T_B}{T} - 1\right)} \tag{2.25}$$

For the pressure control the coordinate scale factor, μ is

$$\mu = \left(1 - \frac{\delta t \beta}{\tau_P}(P_B - P)\right)^{1/3} \tag{2.26}$$

The volume is scaled by the factor μ^3. δt is the time step for the integration of equations of motion[69–71]

2.4.2 Nose-Hoover Thermostat and Barostat

Nose-Hoover method originally proposed by S. Nose and later extended by W.G.Hoover. It is based on the extended Lagrangian as introduced by Anderson consisting of additional thermostating degree of freedom. The external thermal bath coupled to the system is represented by this degree of freedom. The equations of motion of extended system are

$$\dot{\mathbf{R}} = \mathbf{M}^{-1}\mathbf{P} \tag{2.27}$$

$$\dot{\mathbf{P}} = \mathbf{F}(\mathbf{R}) - \frac{p_\eta}{Q}\mathbf{P} \tag{2.28}$$

$$\dot{\eta} = \frac{p_\eta}{Q} \tag{2.29}$$

$$\dot{p}_\eta = \mathbf{P}^T\mathbf{M}^{-1}\mathbf{P} - N_{df}k_BT \tag{2.30}$$

where N_{df} is the number of coordinate degrees of freedom. The parameter Q is the mass of the thermostat (with units of mass times length squared) which determines the size of the coupling. The choice of Q is critical as very large and small values results in costant energy simulation and poor equilibration, respectively. p_η acts like a frictional coefficient either increasing or decreasing the kinetic energy as required to maintain constant temperature.

The Hamilitonian in Nose-Hover algorithm is defined by

$$H = \frac{1}{2}\mathbf{P}^T\mathbf{M}^{-1}\mathbf{P} + V(\mathbf{R}) + \frac{p_\eta^2}{2Q} + N_{df}k_BT\eta \tag{2.31}$$

The simulations performed on the system using the Nose-Hover equations of motion generates trajectories in canonical ensemble. However in few cases it is observed to have poor control over the temperature. This was resolved by the introducion of chain of thermostats.[69–71]

2.4.3 Andersen Thermostat

The temperature control using Andersen algorithm involves the reassignment of the velocities of the particles from Maxwell-Boltzmann distribution during the intervals of normal simulation. Though the trajectories produced are of canonical ensemble, they are discontinous because of the reassignment of velocities.

2.4.4 Langevin Thermostat

The dynamics of the particle interacting with thermal bath is described using stochastic Langevin equation.

$$M\mathbf{a} = F(\mathbf{r}) - \gamma\mathbf{v} + \mathbf{R}(\mathbf{t}) \tag{2.32}$$

Where M, F, **a** represent the mass, force acting and the acceleration of the system, respectively. γ represents the friction coefficient and **R**, the random force. Coupling to the reservior is modeled by adding the fluctuating (R term) and dissipative (-γv term) forces to the Newton equation of motion.

The equation has two extra force terms arising from this interaction - a random force that buffets the particle about and a frictional fore proportional to the particle's velocity, that dissipates excess kinetic energy. In NVE MD simulations, pressure fluctuates more than other quantities such as total energy. The constant pressure is maintained by changing the volume of the system in all directions.[69–71]

2.5 Periodic Boundary Conditions

The goal of the simulations is to study properties of bulk systems (Avogadro's nmber of particles). Periodic boundary conditions (PBC) are applied to the finite-sized simulated system to minimize or eliminate the surface effects in order to obtain error-free bulk properties of the

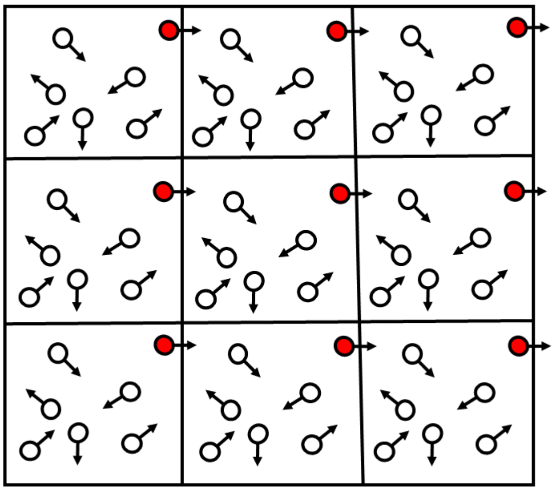

Figure 2.2: Schematic of 2D periodic boundary conditions is shown.

system. Under PBC the primary cell of the particles is replicated periodically by translational operation of the central box along x, y, and z directions as image cells, completely filling the space. Primary and image cells have the same size, shape, number, position and momentum of atoms. A particle leaving the simulation box during the course of simulation is then replaced by an image particle that enters from the opposite side. The particles experience forces as if they were in bulk. As a effect of PBC, the number of interacting pairs increases enormously as each particle in simulation box not only interacts with the other particles in the box but also with their images. This is overcomed with the use of potential with a finite range, all the interactions exceeding the cutoff distance can then be ignored (truncating the interactions beyond a certain cutoff distance). So, among all images of a particle only closest are considered neglecting the rest. This is termed as minimum image convention (criterion).[69–71]

Figure 2.3 shows the steps involved in performing conventional molecular dynamics simulations of system of interest.

2.6 Enhanced Sampling Methods

MD trajectories obtained from tens to hundreds of nanosecond MD simulations inadequately sample the accessible conformational space of the system. Due to Boltzmann sampling, the

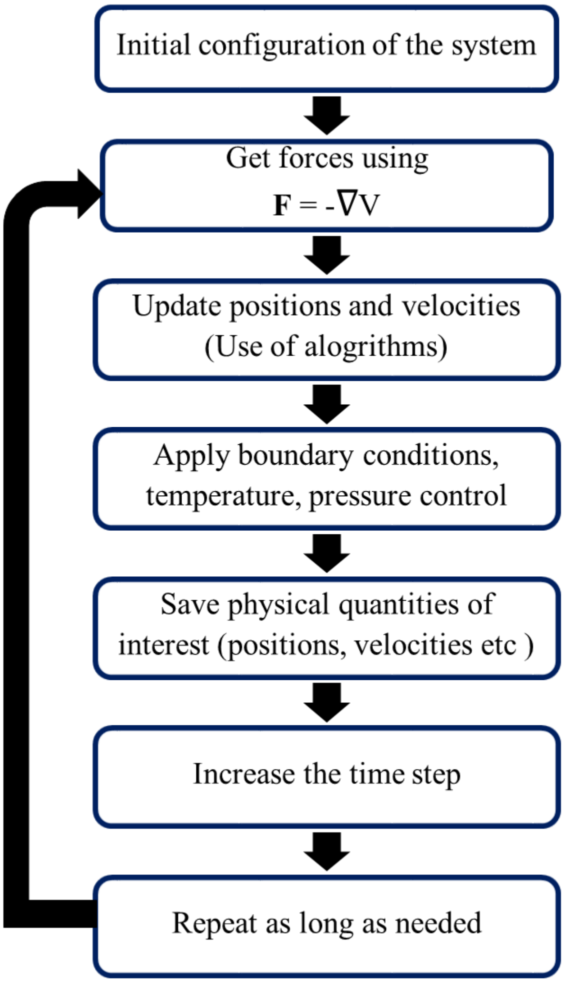
Initial configuration of the system
Get forces using
F = -∇V
Update positions and velocities
(Use of alogrithms)
Apply boundary conditions,
temperature, pressure control
Save physical quantities of
interest (positions, velocities etc)
Increase the time step
Repeat as long as needed

Figure 2.3: Flow chart illustrating the molecular dynamics simulations.

low-energy regions are sampled well while high-energy regions are poorly sampled during the course of the simulation giving rise to non-ergodicity and inaccurate estimates of the equilibrium dynamical and thermodynamic properties.[71,72] Advanced enhanced sampling methods that employ non-Boltzmann sampling are introduced to probe rare events. The primary goal of these methods is to sample efficiently the regions of the conformational space that are important for calculating the free energy of the system. A variety of enhanced sampling methods including umbrella sampling (US), steered MD, metadynamics and adaptive biasing force (ABF) method are used to overcome the conformational sampling problems. In the present thesis, we have employed ABF method to obtain the side chain conformational free energy surfaces of proteins.

2.6.1 Adaptive Biasing Force Method

Consider an N-particle system with a Hamiltonian, H(**r**,**p**)=V(**r**)+T(**p**), where **r** and **p** denote the set of Cartesian coordinates and momenta of all particles, respectively, and V(**r**) and T(**p**) represent the potential and kinetic energies of the system. To determine the potential of mean force, F(ϕ), as a function of a chosen reaction coordinate, ϕ, the Cartesian coordinates are first transformed into a set of generalized coordinates (ϕ, **Q**), where **Q** denotes the set of remaining generalized coordinates. The derivative of F(ϕ) with respect to ϕ is expressed as follows:

$$\frac{dF(\phi)}{d\phi} = \left\langle \frac{\partial V(\phi,\mathbf{Q})}{\partial \phi} - k_B T \frac{\partial \ln |J|}{\partial \phi} \right\rangle_\phi = -\left\langle f_\phi \right\rangle_\phi \tag{2.33}$$

where J is the Jacobian associated with the transformation from the Cartesian coordinates to the generalized coordinates, k$_B$ is the Boltzmann constant, $\left\langle f_\phi \right\rangle_\phi$ is the average force acting along the reaction coordinate, ϕ, determined at a given value of ϕ and the angular brackets denote the statistical averages. In the ABF method, the reaction coordinate ϕ is divided into small windows of size dϕ, $\left\langle f_\phi \right\rangle$ is computed for each bin during the course of the simulation and a biasing force proportional to $\left\langle f_\phi \right\rangle_\phi$ is introduced in the dynamics to generate uniform sampling along the chosen reaction coordinate.[73–76]