CHAPTER 3

METHODS

METHODS

3.1. Computational Techniques:

To study the structural dynamics, aggregation mechanism, polymorphism of amyloid fibrils, and oligomer characterization of the $A\beta_{1-42}$ peptide, we used MD simulation. The principle and theory of the MD simulation is discussed below.

3.1.1. Molecular dynamics simulation:

MD simulations act as a bridge between theory and experiments. As a counterpart to experiment, MD simulations are used to solve scientific problems wherein numerical experiments can be performed for new materials without synthesizing them. MD simulations are used to reproduce experiment to elucidate the invisible microscopic details and to further explain experiments. MD simulation consists of the numerical, step-by-step, solution of the classical equations of motion and thus the molecular interactions can be studied in details. Ultimately to make direct comparisons with experimental measurements made on specific materials a good model of molecular interactions is essential. The crucial advantage of MD simulation lies in its ability to extend the horizon of the complexity that separates 'solvable' from 'unsolvable'.

3.1.1.1. History of simulation:

In the late 1950's MD simulation emerged as one of the first simulation method from the pioneering application to study the interactions of hard spheres by Alder and Wainwright [159]. Many important understandings on the behavior of simple liquids developed from their studies. The next major breakthrough happened in 1964, when the first simulation for liquid argon was carried out using a realistic potential by Rahman [160]. In 1974, Rahman and Stillinger performed MD simulation of liquid water, the first simulation of realistic system like the phase behavior of Lennard-Jones particles [161]. The first protein simulations appeared in 1977 with the simulation of the bovine pancreatic trypsin inhibitor (BPTI) [162]. Due to the revolutionary advances in computer technology and algorithmic improvements, nowadays, MD simulations are used on a regular basis to study the solvated proteins, protein-DNA complexes as well as lipid systems addressing a variety of issues including the thermodynamics of ligand binding and the folding of small proteins. Not only MD simulation but different simulation methods for molecular system have greatly expanded for particular

problems, including mixed quantum mechanical - classical simulations that are being employed to study enzymatic reactions in the context of the full protein. To validate and minimized the experimentally (X-ray crystallography and NMR structure) determined protein structures generally MD simulations are widely used.

3.1.1.2. Theory of molecular dynamics simulation:

The essence of MD simulation consists of integrating Newton's law of motion for a system of interacting particles with mass m and initial positions and velocities with an accurate description of the potential energy as a function of the atomic coordinates. It generates the positions and velocities of the particles in the system that varies with time in phase space and specified as trajectories. These trajectories provide the average values of physical and chemical properties of the particle describing how positions and velocities of the atoms change with time. This is a deterministic method. By solving the differential equation of Newton's second law, the trajectory is attained

$$\vec{F} = ma \quad \text{..... (1)}$$

$$F = -\frac{d}{dr}\mu \quad \text{..... (2)}$$

The forces F is acting on the particles with mass of the particles = m and acceleration of the particle = a, and these are derived from the potential energy μ(r^N), where r^N = (r1, r2 . . . r *N*) represents the complete set of 3*N* atomic coordinates.

The purpose of the numerical integration of Newton's equation of motion is to find an expression that defines position r_i (t+Δt) at time t+Δt in terms of the already known positions at time t. Because of its simplicity, time-reversibility and numerical stability, the Verlet algorithm is frequently used in MD simulation to calculate the trajectories of particles. The basic formula of this algorithm uses Taylor series expansions of the positions and dynamic properties.

A variation on the Verlet algorithm is the leap-frog algorithm [163] where velocities can be calculated from the positions or propagated explicitly.

The leapfrog algorithm use velocities at half time step:

$$\dot{r}_i\left(t + \frac{\Delta t}{2}\right) = \dot{r}_i\left(t - \frac{\Delta t}{2}\right) + \ddot{r}_i(t)\Delta t \quad \text{..... (3)}$$

The velocities at time t can be also computed from:

$$\dot{r}_i(t) = \frac{\dot{r}_i\left(t+\frac{\Delta t}{2}\right)+\dot{r}_i\left(t-\frac{\Delta t}{2}\right)}{2} \quad \text{..... (4)}$$

This is useful when the kinetic energy is needed at time *t*, as for example in the case where velocity rescaling must be carried out. The atomic positions are then obtained from:

$$r_i\ (t+\Delta t) = r_i\ (t) + \dot{r}_i\left(t+\frac{\Delta t}{2}\right)\Delta t \text{..... (5)}$$

The leapfrog algorithm is computationally less expensive and requires less storage. This could be an important advantage in the case of large scale calculations. Moreover, the conservation of energy is respected, even at large time steps. Therefore, the computation time could be greatly decreased when this algorithm is used. However, when more accurate velocities and positions are needed, another algorithm should be implemented, like the Predictor-Corrector algorithm.

The molecular trajectory theoretically imitates the motion of the real system. If the potential energy function is a good estimate of the real interactions between the particles, this can provide an extremely detailed description of both the dynamics and equilibrium properties of the system under consideration. The functional form of the potential energy function together with the set of interaction parameters used is called a *force field*.

3.1.1.3. Force field:

A molecular mechanical force field is an empirical model that describes the interactions within a molecular system. The force field models calculate the energy of a system as a function of the nuclear positions only. Several different force fields have been developed by different research groups. The common thing between all the current force fields is that they use a potential energy function. The typical functional form of a force field is:

$$V(r^N) = \sum_{bonds}\frac{k_i}{2}\left(l_i - l_{i,o}\right)^2 + \sum_{angles}\frac{k_i}{2}\left(\theta_i - \theta_{i,o}\right)^2 + \sum_{torsions}\frac{V_n}{2}\left(1+\cos(n\emptyset-\emptyset_o)\right) + \sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_o\varepsilon_r r_{ij}}\right) \text{..... (6)}$$

Where,

$V(r^N)$: potential energy as a function of the positions (r) of N atoms;
k_i : force constant;
l ,l_0 : current and reference bond lengths;
θ, θ_0 : current and reference valence angle:
V_n : barrier height of rotation;
$Ø$: torsion angle;
n : multiplicity that determines the number of energy minima during a full rotation;
σ_{ij} : collision diameter for the interaction between two atoms *i* and *j*;
ε_{ij} : well depth of the Lennard-Jones potential for the *i-j* interaction;
q_i, qj : partial atomic charges on the atoms *i* and *j*;
r_{ij} : current distance between the atoms *i* and *j*;
ε_0, ε_r : permittivity of the vacuum and relative permittivity of the environment respectively;
$Ø_0$: phase factor that determines where the torsion angle passes through its energy minima.

The potential energy function is generally composed of bonded interactions such as bond lengths, angles and bond rotations and non-bonded interactions i.e. van der Waals and electrostatic interactions. The types of interactions are schematically presented in **Figure 3.1.**

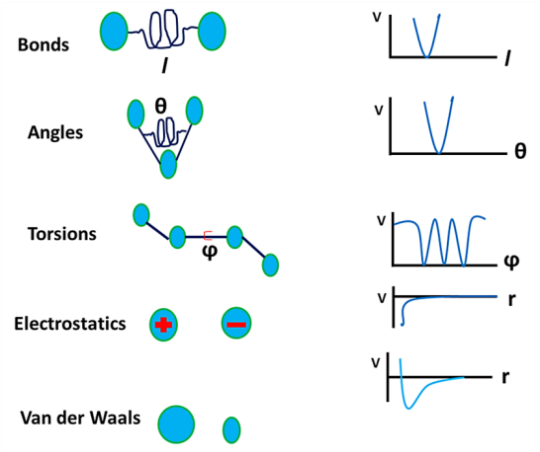


Figure 3.1. *Schematic illustration of the main contribution to the empirical potential energy function (Taken from [164]).*

The first term in the equation represents the bond stretching between pairs of covalently bonded atoms. The second term characterizes the contribution of each angle. Angle bending due to vibrational motions requires less energy to distort an angle from its equilibrium value. The third term models the torsion angle. It shows how the energy changes due to the rotation around a bond. The fourth term of the equation models the

contribution of non-bonded interactions using a Lennard-Jones potential for van der Waals interaction and a Coulomb potential for electrostatic interactions.

3.1.1.4. Long range interactions: Ewald sum:

Ewald summation [165] is one of the most widely used techniques to treat long range interaction in periodic system. Main idea of Ewald sum is to consider a charge distribution of opposite sign on every charge site; this extra charge distribution screens the interaction between neighboring atoms. This screened interaction is now short-range and can be accurately taken care of using the cut-off scheme discussed above for short range interaction. To compensate the additional charge distribution equal charge distribution having opposite sign short range interaction is added and summed in reciprocal space. The contribution to the electrostatic potential at point r_i due to a set of screened charges can easily be computed by direct summation because the electrostatic potential due to screened charge is a rapidly decaying function of r. Total potential energy due to Long range Coulomb interaction is given by the expression:

$$\mu_c = \mu_q(\alpha) - \mu_{self}(\alpha) + \Delta\mu(\alpha) \quad \text{..... (7)}$$

Larger the value of α, sharper the distribution hence large number of K summation has to be included for better accuracy. On the other hand, large value of α reduces range of screened potential hence we can use smaller cutoff radius. Hence value of α is optimized between these two factors to give better accuracy and efficiency. Note that Ewald summation as presented above scales as N 2 only. However, with suitable choice of α and k-space summation cut-off K, Finchman [166] was able to optimize the summation which scales as $N^{3/2}$. Ewald summation can further be optimized through the use of Fast Fourier transform (FFT) in evaluating the reciprocal summation. Particle Mesh based approaches rely on the use of fixed cutoff on the direct space sum together with an FFT based approximation for the reciprocal space sum that scales as *N log(N).*

3.1.1.5. Dealing with molecules: SHAKE algorithm:

In a molecular system, the choice of time step is limited by the various time scales associated with vibrational degrees of freedom such as bond vibration, angle stretching or torsional mode. In general, the bonds involving hydrogen atoms have the fastest vibrational mode and they limit the time step of integration to 1 fs. In order to use a larger time step one can restrain these fast degrees of freedoms while solving the un-

constrained degrees of freedom. Bonds involving hydrogen have the highest frequency hence they can be constrained during dynamics using The SHAKE algorithm which was introduced by Ryckaert *et al* [167]. Basic idea of SHAKE is to use Lagrange multiplier formalism to enforce bonds distances constant. Suppose we have *N*c such constrained given by

$$\propto_k = r^2_{k_1\,k2} - R^2_{k1\,k2} = 0, \text{ where } k = 1, 2, 3\ldots\ldots N\text{c} \quad \ldots\ldots\ldots\ldots\ldots\ldots (8)$$

R_{k1k2} being constrained distant between atoms k_1 and k_2 atoms. This leads to modified constrained equation of motion

$$m_i \frac{d^2 r_i\,(t)}{dt^2} = -\frac{\partial}{\partial_{ri}}\left[V\,(r_1 \ldots\ldots r_N) + \sum_{k=1}^{N_c} \tau_k\,(t)\alpha_k(r_1 \ldots . r_N)\right] \quad \ldots\ldots\ldots\ldots\ldots (9)$$

Where m_i is mass of i^{th} particle and τ_k is the Lagrange multiplier (unknown) for k^{th} constraint. This equation can be solved for unknown multiplier by solving N_c quadratic coupled equations, and we get the following equation of motion:

$$r_{k1}(t + \Delta t) = r^{uc}_{k1}\,(t + \Delta t) - 2(\Delta t)^2 m^{-1}_{k1}\,\tau_k\,(t) r_{k1k2}(t) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (10)$$

Where r_{uc} is position updates with unconstrained force only. This procedure is repeated till defined tolerance is given.

3.1.1.6. Periodic boundary conditions:

The size of the model systems consists of a small number of particles compared to real macroscopic systems. Many of the atoms will experience a large boundary surface to a vacuum environment while simulating which will be irrelevant to study phenomena taking place in bulk. Periodic boundary conditions make it possible to small particles to experience forces if they are in a bulk solution. The atoms are placed in a simulation box that is surrounded by translated copies of the coordinates of the atom as shown in **Figure 3.2**. A periodic 3-dimensional array surrounds the inner cell. If an atom crosses the boundary it is replaced by an image atom that enters from the opposite side with unchanged velocity. Thus, the number of particles within the central box remains constant. A non-bonded cutoff is used to deal with the non-bonded interactions such that each atom interacts with only one image of every other atom in the system.

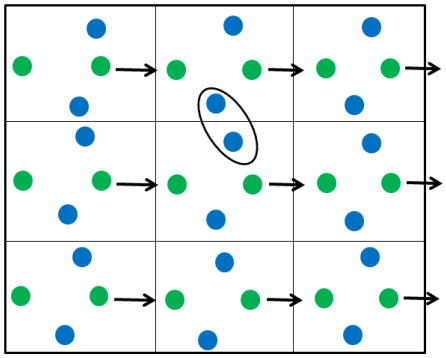


Figure 3.2. *Periodic boundary conditions in two dimensions. The simulation cell (solid) is surrounded by translated copies of itself (dashed).*

3.1.1.7. Temperature and pressure computation and control:

The initial temperature of the system is computed by coupling to a Berendsen thermal bath [168]. The bath supply or remove heat from the system as appropriate, thereby acts as a source of thermal energy. The system temperature T (t) that deviates from the bath temperature T_0 is corrected according to:

$$\frac{dT(t)}{dt} = \frac{1}{\tau}\{To - T(t)\} \quad \text{..... (11)}$$

where τ (time constant) determines the strength of the coupling between the bath and the system. The temperature of the system is corrected by scaling the atom velocities at each step by a factor χ, given by:

$$\chi = [1 + \frac{\Delta t}{\tau_T}(\frac{To}{T(t)} - 1)] \quad \text{..... (12)}$$

The strength of the coupling can be varied by changing the time constantτ.

The method used for pressure control is similar to that of temperature control. The system can be coupled to a barostat and the pressure can be maintained at a constant value by periodic scaling of the simulation cell size and atomic positions with a factor μ:

$$\mu = 1 - \omega\frac{\Delta t}{\tau_p}(\text{P-P}_0) \quad \text{..... (13)}$$

where ω represents the isothermal compressibility, τ_p represents the relaxation constant, P_0 is the pressure of the barostat, P, the momentary pressure at time t and Δt is the time of step. The standard simulation package AMBER12 is used in the present work [169]. *Pmemd*, one of the AMBER modules carries out the molecular dynamics simulation. The various steps involved in setting up and running a MD simulation are discussed below in details and shown in the form of flowchart **Figure 3.3.**

3.1.1.8. Water molecule models:

Many molecular models have been proposed for describing water in MD simulation. These models can be categorized according to the number of sites, the structure (rigid or flexible), and the polarization effects. The 3-site models are the most popular one to be used in MD simulations because of the simplicity, reasonable structural and thermodynamic descriptions and computational efficiency. These kinds of models have three interaction sites which correspond to the three atoms of the water molecule. Each atom gets assigned a point charge. Only the oxygen atom has Lennard-Jones parameters for interaction. Some of the popular 3-site models include transferable intermolecular potential three-point (TIP3P) model , simple point charge (SPC) model, extended simple point charge (SPC/E) model, etc. [170]. Most of these models use a rigid geometry matching the known geometry of the water molecule. The simulations in this thesis are carried out using TIP3P water model. The TIP3P water model used here is specified with the O-H bond length (r_{OH}) and H-O-H bond angle (θ_{HOH}) to be 0.9572 Å and 104.52° respectively which are equal to experimental gas-phase values. The simple model for TIP3P water is shown in **Figure 3.4.**

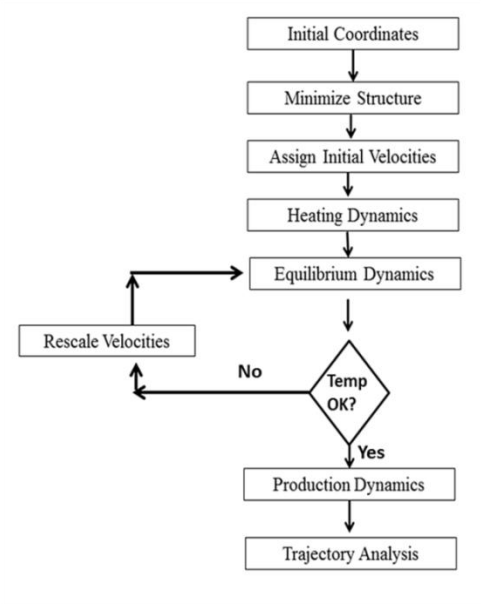


Figure 3.3. *Schematic flowchart of steps involved in MD Simulation.*

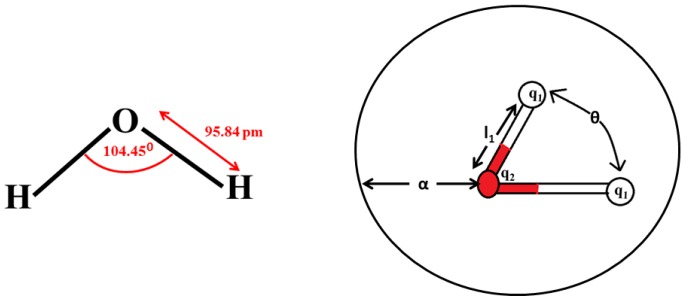


Figure 3.4. *Schematic representation of TIP3P water model.*

3.1.1.9. Molecular dynamics steps:

In order to propagate a molecular system using the above equations there are three typical stages

i. Minimization
ii. Equilibration
iii. Dynamics

(i) **Minimization:**

Using the force field that has been assigned to the atoms in the system it is essential to find a stable point or a minimum on the potential energy surface in order to begin dynamics. At a minimum on the potential energy surface the net force on each atom vanishes. Constraints may be imposed during minimization, as well as during dynamics. These constraints may be based on data such as NOEs from an NMR experiment or they may be imposed by a template such that we force a ligand to find the minimum closest in structure to a target molecule. To minimize we need a function (provided by the force field) and a starting guess or set of coordinates. The magnitude of the first derivative can be used to determine the direction and magnitude of a step (i.e. change in the coordinates) required to approach a minimum configuration. The magnitude of the first derivative is also a rigorous way to characterize convergence.

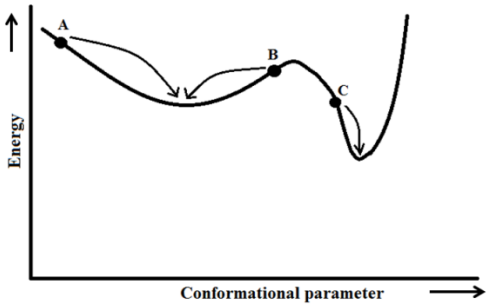


Figure 3.5. *A schematic one-dimensional energy surface. Minimisation methods move downhill to the nearest minimum.*

Most minimization algorithms can go downhill on the energy surface and so they can locate the minimum that is nearest to the starting point. Thus, **Figure 3.5** shows a schematic energy surface and the minima that would be obtained starting from three points A, B and C. To locate more than one minimum or to locate the global energy minimum it requires generating different starting points, each of which is then minimized [171].

A minimum energy is converged when the derivatives are close to zero. Prior to starting a MD simulation, it is important to perform energy minimization of the structure in order to remove the bad contacts, which may otherwise lead to structural distortion. There are three major protocols for minimization:

A. Steepest descent
B. Conjugate gradient
C. Newton-Raphson

A. The Steepest Descents Method:

The steepest descent method determines the direction towards the minimum using the first derivative. It moves in the direction parallel to the net force. For $3N$ Cartesian coordinates this direction is most conveniently represented by a $3N$-dimensional unit vector, $\boldsymbol{s}_k$. Thus:

$$\boldsymbol{s}_k = -\boldsymbol{g}_k/|\boldsymbol{g}_k| \quad \text{..... (14)}$$

Having defined the direction along which to move it is then necessary to decide how far to move along the gradient. Consider the two-dimensional energy surface of **Figure 3.6.** The gradient direction from the starting point is along the line indicated if

we imagine a cross-section through the surface along the line; the function will pass through a minimum and then increase, as shown in the figure. We can choose to locate the minimum point by performing a line search or we can take a step of arbitrary size along the direction of the force [171].

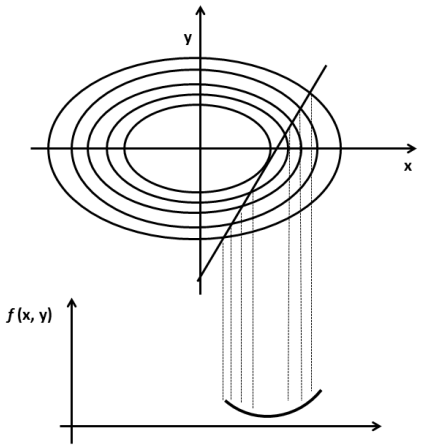


Figure 3.6. *A line search is used to locate the minimum in the function in the direction of the gradient(Adapted from [171]) .*

B. Conjugate Gradients Minimization:

The conjugate method produces a set of directions which does not show the oscillatory behavior of the steepest descents method in narrow valleys. In conjugate gradients, the gradients at each point are orthogonal but the directions are conjugate. A set of conjugate directions has the property that for a quadratic function of M variables, the minimum will be reached in M steps. The conjugate gradients method moves in a direction $\boldsymbol{v}_k$ from point $\boldsymbol{x}_k$ where $\boldsymbol{v}_k$ is computed from the gradient at the point and the previous direction vector v_{k-1} [171].

C. Newton-Raphson Method:

The Newton-Raphson method uses the second derivatives as well as the first derivatives. In addition to using the gradient information, it uses the curvature to predict where along the gradient of the function will change direction. It is the most computationally expensive method utilized to perform energy minimization.

Prior to minimization, water molecules are added to solvate the system if required. A suitable large box of water that has already been equilibrated is used for

solvation purpose. The system is entirely covered by the water box and those water molecules that overlap the proteins are removed. At this point energy minimization should be done with the protein fixed in its energy minimized position. This allows the water molecules to readjust to the protein molecule.

(ii) **Equilibration:**

Molecular dynamics solves the equations of motion for a system of atoms. The solution for the equations of motion of a molecule represents the time evolution of the molecular motions, the trajectory. Depending on the temperature at which a simulation is run MD allows barrier crossing and exploration of multiple configurations. In order to initiate MD we need to assign velocities initially. This is done using a random number generator using the constraint of the Maxwell-Boltzmann distribution. The temperature is defined by the average kinetic energy of the system according to the kinetic theory of gases. The internal energy of the system is U = 3/2 NkT. The kinetic energy is U = 1/2 Nmv^2. By averaging over the velocities of all of the atoms in the system the temperature can be estimated. It is assumed that once an initial set of velocities has been generated the Maxwell-Boltzmann distribution will be maintained throughout the simulation.

Following minimization we can consider the temperature as being essential zero Kelvin. To initialize dynamics the system must be brought up to the temperature of interest. This is done by assigning velocities at some low temperature and then running dynamics according to the equations of motion. After a number of iterations of dynamics the temperature is scaled upwards. The most common means of temperature scaling is velocity scaling. Given a typical time step of 1 fs equilibration is run for at least 5 ps (5000 time steps) and often for 10 or 20 ps.

(iii) **Production dynamics:**

The dynamics stage is the stage of interest for determining thermodynamic averages or sampling new configurations. The stage used for these applications is sometimes known as production dynamics. During this stage of MD simulation thermodynamic parameters can be calculated. Production run can be generated from some hundred ps-ns or more.

3.1.2. Potential of mean force:

The potential of mean force (PMF) [172] is a central concept about the free energy changes as a function of some inter or intramolecular coordinates of molecular systems. The reaction coordinate may be the distance between two atoms, or the torsion

angle of a bond thus fundamentally related to that coordinate's distribution function. The PMF incorporates solvent effects along with the intrinsic interaction between the two particles when the system is in a solvent. The transition state for the process is related to the point of highest energy on the free energy profile, from which rate constant can be derived. There exist various methods to calculate the PMF. The simplest type of PMF is the free energy change with reaction coordinate as the change in separation (r) between two particles and can be defined as [171]

$$A(r) = -k_B T \ln g(r) + \text{constant} \quad \text{..... (15)}$$

The PMF may vary by several multiples of $k_B T$ over the relevant range of the parameter r. The logarithmic relationship between the PMF and radial distribution function means that a relatively small change in the free energy may correspond to g(r) changing by an order of magnitude from its most likely value. Unfortunately, MD simulation method does not adequately sample regions where the radial distribution function differs drastically from the most likely value, leading to inaccurate values from the PMF. One of the most widely used sampling techniques to avoid this problem is the *umbrella sampling (US).*

3.1.2.1. Umbrella sampling:

US overcome the sampling problem by restraining a system to a specific region of its conformational space thereby modifying the potential function so that the unfavorable states are sampled appropriately. The modification of the potential function can be written as:

$$\vartheta'(r^N) = \vartheta(r^N) + W(r^N) \quad \text{..... (16)}$$

Where $W(r^N)$ is a weighting function, which takes a quadratic form:

$$W(r^N) = k_W (r^N - r_0{}^N)^2 \quad \text{..... (17)}$$

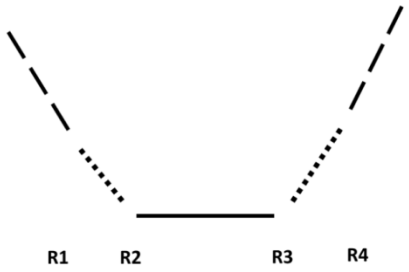
For configurations that are far from equilibrium state $r_0{}^N$ the weighting function will be large and so a simulation using the modified energy function $\vartheta'(r^N)$ will be biased along some relevant 'reaction coordinate' (RC) away from the configuration $r_0{}^N$. The resulting distribution will, of course, be non-Boltzmann. The corresponding Boltzmann averages can be extracted from the non-Boltzmann distribution using a method introduced by Torrie and Valleau [173]. The result is:

$$< A > = \frac{< A(r^N) \exp[+W \frac{r^N}{k_B T}] >_W}{< \exp[+\frac{W(r^N)}{k_B T}] >_W} \quad \text{..... (18)}$$

The subscript *W* indicates that the average is based on the probability P_W (r^N), which in turn is determined by the modified energy function $\vartheta'(r^N)$. It is usual to perform an umbrella sampling calculation in a series of stages, each of which is characterized by a particular value of the coordinate and an appropriate value of the forcing potential W (r^N). However, if the forcing potential is too large, the denominator in Equation 18 is dominated by contributions from only a few configurations with especially large values of exp [W (r^N)] and the average takes too long to converge.

3.1.2.2. Running the umbrella sampling calculations:

With a relaxed starting structure one can run MD on the individual umbrella windows. The key point to remember when selecting the number of windows is that the end points must overlap, i.e. window 1 must sample some of window 2 etc. The force constant similarly has to be big enough to ensure that the subset of phase space are sampled but not too strong that the windows become too narrow and can't overlap.



"\" = lower bound linear response region

"/" = lower bound linear response region

"…" = parabola

"_" = flat region

Normally one can vary the size of the windows and the constraints as a function of position along the pathway. The amount of simulation we do in each window needs to be such we can converge our sampling. To specify the harmonic restraint a reference file is employed where R1, R2, R3, R4 define a flat-welled parabola which becomes linear beyond a specified distance. Essentially between r1 and r2 it will be harmonic

with force constant rk2, between r2 and r3 it will be flat and between r3 and r4 it will be harmonic with force constant rk3.

3.1.2.3. The Weighted Histogram Analysis Method (WHAM) for free-energy calculations:

The WHAM method [174] is an extension of the US method but it has a number of advantages over the conventional US method. The WHAM method, in addition to optimizing the links between simulations, also allows multiple overlaps of probability distributions for obtaining better estimates of the free-energy differences. The older method of obtaining a single distribution function by requiring that the probability distributions agree at some point in the overlap region will fail to yield unique free-energies if three or more distributions are involved in the overlap region. This algorithm provides a built-in estimate of errors that give investigators objective estimates of the optimal location and length of additional simulations to improve the accuracy of their results. The WHAM method takes into account all the simulations that produce overlapping distributions. The WHAM method links the different simulations through the overlapping histograms in an optimal manner. The WHAM equations can also be readily used to generate PMFs and free energies as a function of the coupling parameter(s) *hi* and/or the temperature. This is useful as simulations can be carried out at a range of temperatures to improve conformational sampling and the results extrapolated (or interpolated) to the desired temperature [174].

3.1.3. PatchDock:

PatchDock carries out rigid docking of molecules, whether protein–protein or protein–drug interaction, with surface variability which is addressed through intermolecular penetration [175]. It is based on local shape feature matching algorithm established by Kuntz [176]. At first, the docking method detects the molecular surface areas with a high probability in the binding site, while retaining the correct conformation. The algorithm which they use treats receptors and ligands of variable sizes and thus succeeds in docking of large proteins with small drug molecules. The algorithm functions through three major stages:

(i) **Molecular Shape Representation**: In the first stage two types of surfaces for each molecule are computed. The calculated surface is then preprocessed into distance transform grid and multi-resolution surface which are later used in the

scoring routines. Furthermore, the distance transform grid is used in the shape representation stage of the algorithm. Next, a sparse surface representation is computed which is divided to patches of almost equal area of three types according to their shape geometry: concavities, convexities and flats. Next, the patches are subjected to filter; patches with high propensity to the hot spot residues are retained.

(ii) **Surface Patch Matching**: A hybrid of the Geometric Hashing and Pose-Clustering matching techniques is applied to match the critical points within the patches detected in the previous step. Concave patches are matched with convex and flat patches with any type of patches. Two techniques are used for matching the patches:

A. Single Patch Matching: This type of matching is used for docking of small ligands, like drugs or peptides, wherein one patch from the receptor is matched with one patch from the ligand.

B. ***Patch-Pair Matching***: This type of matching is used for protein-protein docking, wherein two patches from the receptor are matched with two patches from the ligand.

*(iii)***Filtering and Scoring**: Unacceptable steric clashes between the receptor and ligand atoms are discarded in the last stage and score is given.

A. Steric Clashes Test: In this stage the distance transform grid is extensively used. The transformation is applied on the surface points of the ligand. Next the distance transform grid of the receptor is matched with the coordinates of every surface point. If the distance is less than penetration threshold for each surface point, the transformation is retained for the next step, otherwise the transformation is disqualified.

B. Geometric Scoring: The general idea is to divide the receptor into shells according to the distance from the molecular surface. Each shell is defined by a range of distances in the distance transform grid. The geometric score is a weighted average of all the shells, where candidate complexes with large number of points in the shell, and as little as possible points in the ’penetrating’ shells are preferred.

3.1.4. The molecular mechanics energies combined with the Poisson-Boltzmann or generalized Born and surface area continuum solvation method (MM-PBSA and MM-GBSA):

The MM-PBSA and MM-GBSA methods was originally defined by Kollman *et al.* [177] are characterized by the use of Poisson-Boltzmann (PB) and Generalized Born (GB) models to compute the absolute binding free energy for the non-covalent association of any two molecules, A and B, in solution, that is.

$$[A]_{aq} + [B]_{aq} \Leftrightarrow [A^* B^*]_{aq^*} \quad \text{..... (19)}$$

where $[A]_{aq}$ refers to the dynamical structure of molecule A free in solution, $[B]_{aq}$ refers to the dynamical structure of molecule B free in solution, and $[A^* B^*]_{aq^*}$ represents the complex formed from molecules A and B. The binding free energy for the noncovalent association of two molecules may be written in terms of thermodynamic quantities as:

$$\Delta G = \Delta H - T\Delta S \quad \text{..... (20)}$$

Wherein, ΔH is the enthalpy, ΔS represents entropy and T is the temperature of the system at 300 Kelvin. The binding free energy (ΔG) of a receptor-ligand complex is computed as:

$$\Delta G_{bind} = G_{com} - [G_{rec} + G_{lig}] \quad \text{..... (21)}$$

Gcom is the absolute free energy of the complex, Grec is the absolute free energy of the receptor, and Glig is the absolute free energy of the ligand. The enthalpy term in equation 20 can be dissected into sub-energy terms:

$$H_{tot} = H_{gas} + G_{solv} \quad \text{..... (22)}$$

$$H_{gas} = E_{el} + E_{vdW} + E_{int} \quad \text{..... (23)}$$

Hgas is the potential energy of the solute which is determined as the sum of van der Waals (Evdw), electrostatic (Eel) and internal energies (Eint) in gas phase. Gsolv is the solvation free energy for transferring the solute from vacuum into solvent and is a sum of electrostatic (Gel) and non-electrostatic (hydrophobic) contributions (Gnonel) as shown in equation 24:

$$G_{solv} = G_{el} + G_{nonel} \quad \text{..... (24)}$$

The total entropy (Stot), as formulated in equation 25 arose from changes in the degree of freedom:

$$S_{tot} = S_{trans} + S_{rot} + S_{vib} \quad \text{..... (25)}$$

In equation 25, (*S*trans) is the translational, (*S*rot) the rotational, and (*S*vib) the vibrational entropy of each species. Considering all absolute energy terms as given in equation 21, the binding free energy ΔG takes the following form:

$$\Delta G_{\text{binding}} = [\Delta H_{\text{gas}} + \Delta G_{\text{solv}}] - T\Delta S_{\text{tot}} \quad \text{..... (26)}$$

3.1.5. Contact Map Analysis:

For a given PDB file, the 'Contact Map Analysis' server (CMA) evaluates residue–residue contacts between two chains or within a single one [178]. The interface contacts between the two chains in a given PDB file is considered and residue–residue contacts are represented as an interactive contact map. The CMA server is based on LPC/CSU software which gives in detail information including names of the contacting atoms, distances and atom–atom contact areas. The program evaluates simultaneously contact surface area and solvent-accessible surface area. The analysis of ligand-protein interaction is based upon an approach termed surface complementarity [179]. The complementarity function is defined as:

$$\text{CF} = S_1 - S_i - E \quad \text{..... (27)}$$

Where S_1 indicates the sum of all surface areas of 'legitimate' atomic contacts between ligand and receptor, S_i indicates the surface areas of 'illegitimate' atomic contacts and E is a repulsive term. Legitimacy depends upon the hydrophobic/hydrophilic properties of the contacting atoms. As input, LPC software reads PDB formatted file consisting of coordinates of protein atoms and ligand(s). It then assigns atom class to every atom of protein and ligand based on legitimacy and interatomic distances. Two atoms are considered to be covalently bound if the distance between them is < 2.0 Å.

3.1.6. PDBsum:

PDBsum is a Web-based database [180]. The database provides a pictorial summary on each macromolecular structure deposited at the Protein Data Bank together with various analyses of their structural features. PDBsum server also provides a schematic depiction of the inter-molecular interactions as a LIGPLOT diagram [181]. The LIGPLOT program automatically generates schematic 2-D representations of protein-ligand complexes from standard Protein Data Bank file input. The output is a color, or black-and-white, PostScript file giving a simple and informative representation of the intermolecular interactions and their strengths, including hydrogen bonds, hydrophobic interactions and atom accessibilities. The program is completely general

for any ligand and can also be used to show other types of interaction in proteins and nucleic acids [182].

3.1.7. Analysis of trajectories:

(i) Root Mean Square Deviation (RMSD): The deviation of a structure with respect to a particular conformation is measured by RMSD. It is defined as:

$$\text{RMSD} = \left(\frac{\sum_N \left(R_i - {R_i}^0\right)^2}{N}\right)^{1/2} \quad \text{.....(28)}$$

where N is the total number of atoms/residues considered in the calculation, and Ri is for the vector position of particle i (target atom) in the snapshot, R_i^0 is the coordinate vector for reference atom i. RMSD was computed based on backbone atoms and taking the first frame of the simulation as the reference.

(ii) Root Mean Square Fluctuation (RMSF): It is useful for characterizing local changes along the protein chain. It is calculated as:

$$\text{RMSF} = \left(\frac{1}{T}\sum_{t=1}^{T}\left(r_i(t) - {r_i}^{ref}\right)^2\right)^{1/2} \quad \text{..... (29)}$$

T is the trajectory time over which the average is taken, r_i (t) is the position of the atoms in residue *i* and r_i^{ref} is the reference position of particle *i*.

(iii) Radius of Gyration (Rg): It calculates the distribution of the components of an object around the axis. It gives the compactness of a protein. It is calculated as:

$$\text{Rg} = \left(\frac{1}{N}\sum_i (r_i - r_{cm})^2\right)^{1/2} \quad \text{..... (30)}$$

where r_i - r_{cm} is the distance between atom i and the center of mass of the molecule.

(iv) Secondary Structure Analysis: The secondary structure content of each protein was calculated using DSSP algorithm which recognizes cooperative secondary structures as repeats of the elementary hydrogen-bonding patterns "turn" and "bridge."

Repeating turns are “helices,” repeating bridges are “ladders,” connected ladders are “sheets. We consider that a secondary structure element is stable at a given position of the protein if it is the predominant in > 50% of the collected snapshots [183].