CHAPTER 4

Material and Methods

4. Materials and Methods:

4.1. Tools and techniques:

Several experimental methodologies have been utilized to assess the effects of point mutations on the structure and functionality of SHANK3. Molecular dynamics (MD) simulation has been our approach to investigate the stability, flexibility, and intramolecular interactions within the N-terminal region of SHANK3. Moreover, R programming is utilized to explore the significant differentially expressed genes in ASD samples and to identify potential biomarkers for early detection. The fundamental procedures involved in MD simulation and R programming packages are elaborated upon below:

4.1.1. Molecular Dynamics (MD) Simulation:

Proteins constitute essential components of living organisms, playing a key role in critical cellular processes necessary for life, including molecular recognition, signal transduction, protein localization, and enzyme catalysis [1]. These biological activities are governed by protein movements and physical interactions with other molecules like ligands, peptides, proteins, and nucleic acids. Recent advancements in structural biology and biophysical characterization techniques, such as X-ray crystallography, nuclear magnetic resonance (NMR), and cryo-electron microscopy, have led to a vast expansion in the number of available three-dimensional structures for proteins, protein-ligand complexes, and protein-protein complexes. However, observing the dynamic behavior of individual atoms and manipulating them at the atomic level poses significant challenges. An appealing alternative is employing atomic-level computer simulations of relevant biomolecules [2], facilitating the study of protein dynamics at an atomic scale.

The introduction of molecular dynamics (MD) simulation to molecular biology has revolutionized researchers' ability to visualize microscopic biological processes [3-5]. Molecular dynamics simulations forecast the movement of every atom within a protein or molecular system over time based on a comprehensive model of the physics governing interatomic interactions [5]. These simulations capture a broad spectrum of crucial biomolecular phenomena, including conformational changes, ligand binding, protein stability, and protein folding, providing detailed atomic positions with femtosecond temporal resolution. Importantly, such simulations predict how biomolecules will react—at an atomic level—to various perturbations like mutations,

phosphorylation, protonation, or the addition/removal of ligands. When integrated into the drug discovery pipeline, these applications empower researchers to identify essential interactions required for the favorable binding of small molecules, peptides, and proteins to binding pockets or protein-protein interfaces.

4.1.1.1. History of simulation:

Alder and Wainwright are credited as the forerunners of molecular dynamics (MD) simulation, having pioneered this methodology in the late 1950s to investigate the interactions of hard spheres [6, 7]. In 1964, Rahman conducted the initial simulation utilizing a realistic potential for liquid argon [8]. Subsequently, the first simulation of a realistic system, specifically liquid water, was accomplished in the 1970s [9]. Furthermore, the initial simulation of proteins was applied in 1977 [3]. Subsequently, during the 1980s, advancements were made in simulating protein interactions with small molecules, elucidating their thermodynamics through free energy calculations, and rapid computation of biomolecules [10]. In 1998, Duan and Kollman demonstrated the folding mechanism of a small sub-domain of villin through a microsecond-scale simulation [11]. Notably, in 2013, Martin Karplus, Michael Levitt, and Arieh Warshel were awarded the Nobel Prize in Chemistry for their contributions to the development of multi-scale models for complex chemical systems, enabling simulations of molecular behavior across various scales, from individual molecules to proteins. The growing significance of MD simulation has stimulated the development of numerous techniques, including potential sampling methods, advancements in force fields, and the availability of high-performance computational resources, enabling simulations spanning microseconds to milliseconds. MD simulation emerges as a valuable tool for investigating biomolecular dynamics. However, effective utilization of MD simulation necessitates the development of optimal models capable of accurately representing the cellular environment, as well as physical methodologies to induce motion within the model, alongside substantial computational resources. Nevertheless, recent progress has seen the extension of MD simulations to cellular scales, facilitating simulations encompassing entire cells [12]. Further refinement of algorithms and theoretical frameworks for modeling, docking, scoring, and energy calculations promises to enhance the efficacy of MD simulation methodologies.

4.1.1.2. Theory of MD simulation:

The cornerstone of MD simulation hinges upon Newton's law of molecular mechanics [13], which embodies Newton's second law of motion: **F = ma**, where '**F**' represents the force applied on a particle, '**m**' denotes its mass, and '**a**' defines its acceleration. This essential equation underpins the MD simulation methodology. When the forces acting on individual atoms are known, it becomes feasible to compute the acceleration for each atom within the system. Integration of the motion equations generates a trajectory that delineates the positions, velocities, and accelerations of particles over time. From these trajectories, insights into the average values of particle characteristics can be inferred. Notably, this approach is deterministic, allowing for the prediction of system states in past, present, or future time frames once the velocities and positions of all atoms are established.

Despite the deterministic nature of MD simulation, its application can be resource-intensive, necessitating significant time and computational resources. Nonetheless, with the increasing affordability and speed of computers, simulations for solvated proteins have been extended to the millisecond time scale. Furthermore, studies have reported simulations conducted within the millisecond timeframe for various molecular systems.

Newton's law-based equation of motion is as follows: -

$$F_i = m_i a_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.1})$$

$$\vec{F} = ma \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.2})$$

$$F = -\frac{d}{dr}\,\mu \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.3})$$

In this equation, F_i defines the force exerted on particle i, m_i as the mass of particle i, and a_i as the acceleration of particle I, which have been derived from the potential energy $\mu(r^N)$, where r N = (r1, r2 . . . r N) denotes the entire set of 3N atomic coordinates.

The Newton's force, F_i can also be expressed as the potential energy gradient,

$$Fi = -\nabla iV \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.4})$$

The two equations can be combined,

$$-\frac{dV}{dr_i} = m_i \frac{d^2 r_i}{dt^2} (1+x)^n$$

$$= 1 + \frac{nx}{1!} + \frac{n(n-1)x^2}{2!} + \cdots \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.5})$$

In this context, the system's potential energy, denoted as V, is elucidated. A relationship exists between Newton's equation of motion and the derivative of potential energy, which serves to elucidate changes in position over time.

The primary aim of numerically integrating Newton's equation of motion is to creat a formula that delineates the position, r*i* *(t*+Δ*t* at time *t*+Δ*t*, based on the known positions at time t. The *Velocity Verlet* method utilizes positions and accelerations at time t, along with positions from time *t*-Δ*t*, to determine the new positions at time *t*+Δ*t*. Direct computation of velocities is not entailed in this procedure; however, their calculation is essential for determining the kinetic energy, K. Throughout this process, the total energy, E=K+U, must be conserved.

The position of each atom is computed at every Δt time step: -

$$Ri(t+\Delta t) = \frac{ri(t) + vi(t)\Delta t \; + \; 1\, ai(t)\Delta t)^2}{t} \ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.6})$$

The velocity is utilized just as a half time step: -

$$r_i(t + \frac{\Delta t}{2}) = r_i\left(t - \frac{\Delta t}{2}\right) + r_i(t)\Delta t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.7})$$

The velocities can be calculated from the Δt time step: -

$$r_i(t) = \frac{r_i(t + \frac{\Delta t}{2}) + r_i(t - \frac{\Delta t}{2})}{2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.8})$$

It is important to note that velocity rescaling is necessary when kinetic energy is needed at time t. Furthermore, the required atomic sites are acquired from:

$$r_i(t + \Delta t) = r_i(t) + r_i\left(t + \frac{\Delta t}{2}\right)\Delta t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.9})$$

Force fields are employed to depict the alterations in bond angles, bond lengths, torsions, as well as non-bonding van der Waals and electrostatic interactions among atoms over time. Comprising interconnected constants and equations, this force field endeavors to emulate the molecular geometry and distinctive attributes of the examined structures.

4.1.1.3. Force field:

A force field, in the context of molecular modeling, is a mathematical equation utilized to characterize the dependence of the energy system on the positions of its constituent particles. It typically comprises an analytical interatomic potential energy function, denoted as U(*r1, r2,... rN*), along with other contributing factors. The parameters of a force field are often determined through fitting procedures to experimental data obtained from various sources, such as neutron, X-ray, and electron diffraction, NMR, infrared, Raman, and neutron spectroscopy, or alternatively from ab initio or semi-empirical quantum mechanical calculations. Essentially, a force field provides a simplified model of a collection of atoms bound together by basic elastic (harmonic) forces, which is valid within the simulation domain as a substitute for the true potential energy landscape.

$$\begin{aligned} V(r^N) = & \sum bonds\ \frac{k_i}{2}(l_i - l_i, o)^2 + \sum_{angles} \frac{k_i}{2}(\boldsymbol{\Theta}_i - \boldsymbol{\Theta}_i, O)^2 \\ & + \sum torsions \sum_{torsions} \frac{V_n}{2}(1 + \cos(n\emptyset - \emptyset_o)) \\ & + \sum_{i=1}^{N}\sum_{j=i+1}^{N} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \\ & + \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_{ij}}\Bigg) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.10}) \end{aligned}$$

Here,

$V(r^N)$*:* 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 defines the number of energy minima throughout a complete rotation

σ_{ij}: collision diameter for i and j, the two atoms interacting;

q_i, q_j: partial atomic charges on the atoms i and j;

ε_{ij}: well depth of the Lennard-Jones potential for the i-j interaction;

r_{ij}: The present distance between atoms i and j;

ε_0, ε_r: relative permittivity of the environment and the permittivity of the vacuum, respectively;

$Ø_0$: phase factor that determines the torsion angle's energy minimum at every position.

The majority of the potential energy function comprises bond lengths, angles, and rotations, along with non-bonded interactions such as van der Waals and electrostatic interactions. An illustrative depiction of these diverse interactions is provided in **Figure 4.1**.

$$U(R) = \sum_{bonds} k_r (r - r_{eq})^2 \quad \textit{bond}$$

$$+ \sum_{angles} k_\theta (\theta - \theta_{eq})^2 \quad \textit{angle}$$

$$+ \sum_{dihedrals} k_\phi (1 + \cos[n\phi - \gamma]) \quad \textit{dihedral}$$

$$+ \sum_{impropers} k_\omega (\omega - \omega_{eq})^2 \quad \textit{improper}$$

$$+ \sum_{i<j}^{atoms} \varepsilon_{ij} \left[\left(\frac{r_m}{r_{ij}} \right)^{12} - 2 \left(\frac{r_m}{r_{ij}} \right)^{6} \right] \quad \textit{van der Waals}$$

$$+ \sum_{i<j}^{atoms} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \quad \textit{electrostatic}$$

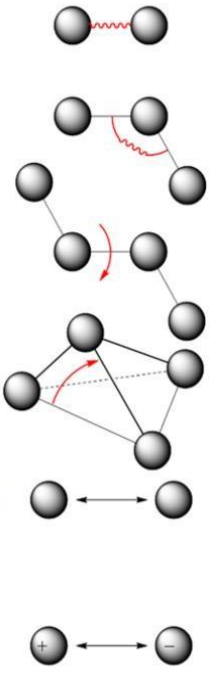


Figure 4.1. *Schematic exhibiting the primary contribution to the potential energy function. (Right) potential energy terms in a force field, and (Left) energy function used*

to derive atomic forces for molecular movement. r is the bond length; θ is the atomic angle; ϕ is the dihedral angle; ω is the improper dihedral angle; r_{ij} is the distance in between atom i and j; k_r, k_θ, k_ϕ, and k_ω are force constants; r_{eq}, θ_{eq}, and ω_{eq} are equilibrium positions; the dihedral term is a periodic term characterized by a force constant (k_ω), multiplicity (n), and phase shift (γ); ε_{ij} is related to the Lennard–Jones well depth; rm is the distance at which the potential reaches its minimum; qi and q_j are the charges on the respective atoms; and ε_0 is the dielectric constant[14].

The initial four terms within the equation denote the local or intramolecular factors influencing the total energy, encompassing aspects. The last two terms of the equation delineate the repulsive van der Waals interactions and Coulombic contacts, with this instance employing a 12-6 Lennard-Jones potential.

4.1.1.4. Long range interaction: Ewald sum:

In computational simulations of condensed-matter systems, the electrostatic interactions are commonly computed employing the Ewald Summation method [15]. The Ewald Sum construction, errors stemming from truncating the infinite real- and Fourier-space lattice sums are tested. An optimal choice is determined for the Fourier-space cutoff, typically set at a screening parameter of 7. However, it is apparent that a certain level of precision is associated with a scaling factor of 7/3, irrespective of the number of vectors encompassed in the Fourier space. Nevertheless, by prevailing the efficient computation parameters for Ewald sums, this proposed methodology can be employed to evaluate the quality of Ewald-sum implementations and compare various implementations. In MD simulations, this methodology may be implemented most frequently to evaluate long-range interactions. The Ewald sum is predicated on the analysis of a charge distribution for the opposite sign of every single site. The additional charge distribution reveals the interactions between all surrounding atoms. The threshold scheme can control the interactions, albeit in their restricted range. To make up for the excess charge distribution, the identical charge distribution with the opposite sign and short-range interaction is built up in the reciprocal space. It is straightforward to retrieve the input for the electrostatic potential at a particular position ri produced by an ensemble of screened charges since the electrostatic potential resulted from the screened charge is a fast diminishing function of r. The aforementioned formula furnishes the total potential energy for the long-range Coulomb interaction:

$$\mu_c = \mu_q(\alpha) - \mu_{self}(\alpha) + \Delta\mu(\alpha) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.11})$$

More sharp distributions are generated by higher values of α in the formula; to boost precision for such large numbers, K summations are added. A higher value for α, on the contrary, minimizes the filtered potential range and permits us to choose a lower threshold radius. Consequently, there is room for optimization of the value of α between the two factors in order to maximize both efficiency and accuracy. The Ewald summation is represented by N2 solely in the aforementioned scales. However, by choosing the right values for α and K for the k-space summation cutoff, Finchman was able to optimize the summing that scales as $N^{(3/2)}$. Additionally, by evaluating the reciprocal summation with the Fast Fourier Transform (FFT), this Ewald summation method can be enhanced. On the other hand, the particle mesh-based solution relies on the usage of an estimated reciprocal space sum based on FFT that grows as N log (N) and a set cutoff for the direct space sum. The infinite-range Coulomb interaction is efficiently calculated by this method under periodic boundary conditions (PBC). Furthermore, there is a version called Particle Mesh Ewald (PME) that accelerates the Ewald reciprocal sum to almost linear scaling by using the three-dimensional fast Fourier transform (3DFFT). Under PBC, particle i inside the unit cell interacts electrostatically with every other particle j inside the cell, with every periodic image of j, and with each of its own periodic pictures because of the infinite range of the Coulombic interaction. The total Coulomb energy of a system consisting of N particles in a size L cubic box and all of their infinite duplicates in PBC is given by **equation 4.12**:

$$U = \frac{1}{2}\sum_{n}^{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{r_{ij,n}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.12})$$

Ewald reformulated a single slowly and conditionally converging series, **equation (4.13)**, as the product of two rapidly convergent series plus a constant term,

$$U_{Ewald} = U^r + U^m + U^0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.13})$$

Therefore, the total of these three elements—the real (direct) space sum (Ur), the reciprocal (imaginary, or Fourier) sum (Um), and the constant term (U°), also known as the self-term denotes the Ewald sum.

4.1.1.5. SHAKE algorithm: Dealing with molecules:

The choice of time step is limited by the many time scales related to vibrational degrees of freedom, such as bond vibration, angle stretching, or torsional modes inside a molecular system. The integration time step is limited to 1 fs due to the bonds involving hydrogen atoms having a faster vibrational state. To necessitate a longer time period, however, one can restrain these rapid degrees of freedom while addressing the unconstrained degrees of freedom. Since hydrogen bonds have the greatest frequency, the SHAKE technique, created by Ryckaert et al., can restrict dynamics for these types of bonds [16]. Relaying the unconstrained equations of motion of the atomic system is the first step in the SHAKE algorithm. The fundamental idea behind the SHAKE method is to use the Lagrange multiplier formalism to enforce bonding distances to remain constant. Given N_c, the constraint is provided by:

$$a_k = r^2{}_{k1\,k2} - R^2{}_{k1\,k2} = 0, where\ k = 1,2,3 \ldots\ldots N_C \ldots\ldots (\mathbf{4.14})$$

The distance between the atoms of k1 and k2 is thought to be limited by the parameters R_{k1k2}. The following defines the modified constrained equation of motion:

$$m_i \frac{d^2 r_i(t)}{dt^2} = -\frac{\partial}{\partial_{ri}}[V\,(r_1 \ldots . r_N) + \sum_{k=1}^{Nc} \tau_k(t)\,\alpha_k(r_1 \ldots\ldots r_N)] \ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.15})$$

In this case, τ_k represents the unknown Lagrange multiplier for the kth constraint, and m_i is the mass of the i^{th} particle. N_c quadratic coupled equations can be used to solve this modified restricted equation of motion for an unknown multiplier. Ultimately, the motion **equation 4.16** that follows has been determined.

$$r_{k1}(t + \Delta t) = r_{k1}^{uc}(t + \Delta t) - 2(\Delta t)^2\, m_{k1}^{-1}\, \tau_k\,(t) r_{k1k2}(t) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.16})$$

The position updates with unconstrained force solely are represented by r^{uc} in the **equation 4.16**. But this method is repeated if the specified tolerance is not provided.

Instead of explicitly inverting the matrix, the SHAKE algorithm approach modifies the particle coordinates periodically until the system satisfies all requirements within a certain tolerance. Constraint decay, or the increase in departure from the ideal lengths caused by the accumulation of numerical errors, is a factor that constraint algorithms must take into consideration in addition to maintaining the rigid bonds. However, since the convergence of each time step must happen, iterative algorithms enable accurate constraint decay automatically within a certain tolerance. The constrained distance deviations from the initial values undergo frequent inspections and fixes. Non-iterative algorithms required a deliberate strategy to counter constraint deterioration since they lacked a natural feedback mechanism for detecting changes in distance.

4.1.1.6. Periodic Boundary conditions:

To explain the periodic boundary conditions, we will construct a system consisting of N interacting particles in a volume V and at a temperature T. The system must be tied by copies of itself in order for us to guarantee periodic limits on boundaries that are similar to the 2D Using system. Consequently, it can be observed that, given a system of particles, a particle must rejoin the central box on the opposite side where it departs. The atoms of the molecules are organized, as shown in **Figure 4.2**, in a hypothetical box bounded by translated copies of their coordinates. Particle 1 inside the middle box may potentially interact with many copies of particle 3 that are present there, as shown in **Figure 4.2**. Furthermore, considering a particular interaction between particles 1 and 3 is suitable, and selecting the interaction that results in the smallest interatomic distance makes sense. This method is referred to as the closest image convention. There is evidence that a periodic 3-dimensional array encircles the inner cell. An atom gets replaced when it passes through a barrier and enters the opposite side at the same speed. The particles in the core box have a fixed volume after this. However, in order to handle non-bonded interactions, a non-bonded cutoff is typically used, allowing each atom in the system to interact with just one image of each and every other atom in the system.

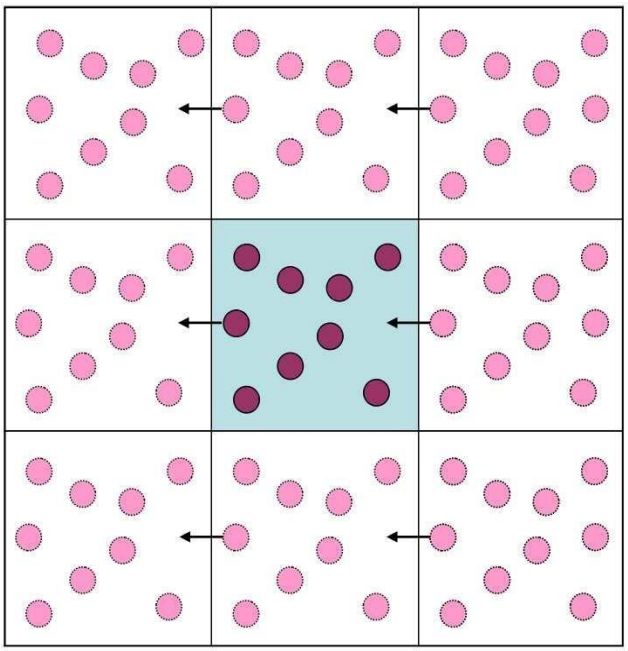


Figure 4.2. *The two dimensional projection of Periodic boundary conditions. The simulation cell (dark color) is surrounded by translated copies of itself (light color).*

4.1.1.7. Temperature and Pressure computation control:

Multiple strategies are being investigated at the moment to achieve the isothermal MD simulations. Adding a fictitious heat bath to the system in order to maintain the average temperature **T** at a certain target temperature $\mathbf{T_0}$ is theoretically equivalent to utilizing a thermostat. However, the heat bath still matters for a particular particle I since it might alter the particle's velocity or modify Newton's equation of motion:

$$m_i \frac{\mathrm{d}^2 r_i}{\mathrm{dt}^2} = F_i(r_i) \quad \ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.17})$$

The system usually evolves according to the above-described **equation 4.17**, which has a micro-canonical (NVE) energy distribution in the absence of temperature control. This micro-canonical ensemble provides the value of "real" dynamics, such as classical Newtonian dynamics, for a system described by an individual force field at the precision level limited by the integration method and the force calculations. By attaching the system to a Berendsen thermal bath, the starting temperature of the system is determined. The bath serves as a source of thermal energy by adding or removing heat from the system as needed. The following **equation 4.18** is used to adjust the system temperature T(t) that differs from the bath temperature T_0:

$$\frac{dT(t)}{dt} = \frac{1}{\tau}\{T_0 - T(t)\} \dots\dots\dots\dots\dots\dots\dots\dots\dots (\mathbf{4.18})$$

Where the intensity of the coupling between the bath and the system is determined by the time constant, or τ. The system's temperature is adjusted by dividing each step's atom velocities by a fraction χ, which is given by **equation 4.19**: A change in the time constant τ can alter the coupling's strength.

$$\chi = \left[1 + \frac{\Delta t}{\tau T}\left[\frac{T_0}{T(t)} - 1\right]\right] \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (\mathbf{4.19})$$

The temperature control method and the pressure control approach are comparable. The system may be connected to a barostat, and by periodically scaling the atomic locations and simulation cell size by μ, the pressure can be kept constant: **equation 4.20**

$$\mu = 1 - \omega\frac{\Delta t}{\tau_p}(P - P_0) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (\mathbf{4.20})$$

Here, ω is the isothermal compressibility, P_0 is the barostat pressure, P is the momentary pressure at time t, Δt is the step time, and $\boldsymbol{\tau}_p$ is the relaxation constant. The AMBER 14 standard simulation software is used. The MD is carried out using the PMEMD one module of AMBER [17].

4.1.1.8. Water molecule model:

Three-site water models were followed by the development of four-site water models. The Bernal and Fowler model is the first of the four-site models. It was created in 1933 and is currently only sometimes used, but it is significant historically [18]. As on the negative charge is moved from the oxygen and towards the hydrogens in that particular model at the bisector of the HOH angle, which is 0.15 Å from the oxygen atom. In either case, the centre of the Lennard Jones interaction site contains the oxygen atom. Ten distances, as opposed to nine, are needed to evaluate the interaction function for a three-site model. Numerous levels of approximation (such as quantum vs classical, flexible versus rigid) and the intricacy of water characteristics have resulted in the suggestion of hundreds of theoretical and computational models for water [19]. Among the classical water models, the rigid non-polarizable models that represent water as a collection of point charges at fixed locations relative to the oxygen nucleus are the most

straightforward and computationally efficient. These models are the class that is employed in the great majority of biomolecular investigations conducted today. The most often used models in this class, such as the triple point charge (TIP3P) [20] (as shown in **Figure 4.3**) and SPCE [21] 3-point models, the four point charge ($TIP4P_{ew}$) [22] 4-point model, and the TIP5P [23] 5-point model, have generally succeeded in striking a fair balance between speed and accuracy, albeit they are far from ideal. The basic TIP4P water model is re-parameterized for use with Ewald methods, offering a worldwide increase in water characteristics overall when compared to a number of widely-used polarizable and non-polarizable water potentials. The new $TIP4P_{ew}$ potential has a density maximum at approximately 1°C and reproduces experimental bulk densities and the enthalpy of vaporization with an absolute average error of less than 1%, from -37.5 to 127 °C at 1 atm, using high precision simulations and careful application of standard analytical corrections [23].

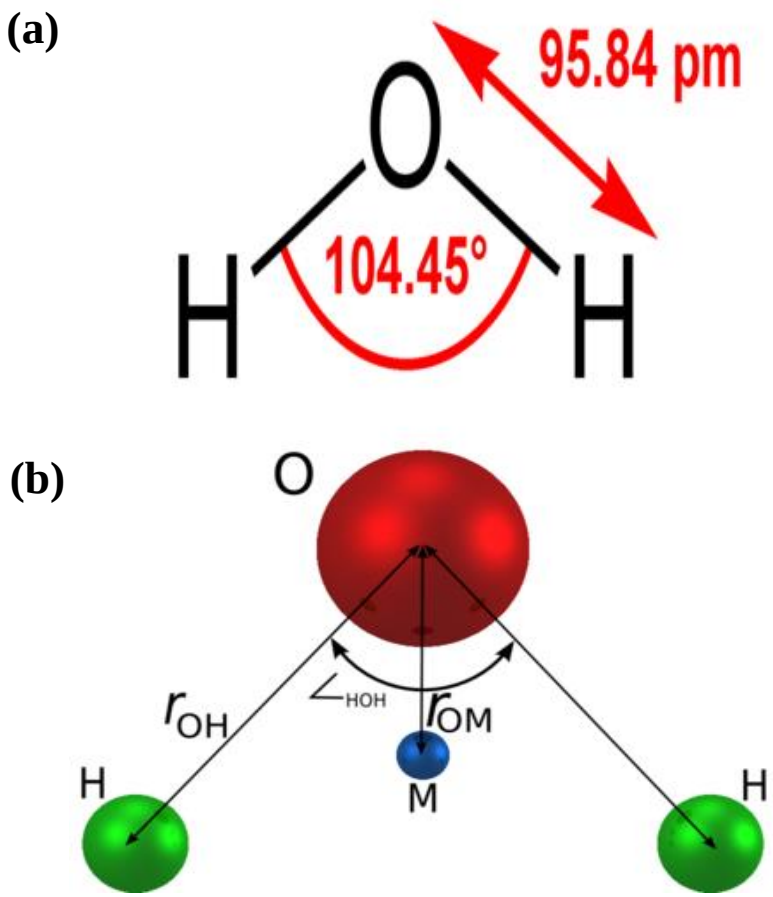


***Figure 4.3**. Schematic representation of TIP3P water model (Adapted from [23])*

It takes 10 distances instead of nine to evaluate the interaction function for a three-site model. As an example, the ST2 model requires 17 participants in a five-site strategy.

The TIPS2 and TIP4P potentials were later developed by Jorgensen and associates using just the BF water model as a basis [24]. A negative charge is further provided to a site located on the bisector of the HOH angle in a manner similar to the TIP4P water model. Both the hydrogen sites and the oxygen are impacted by the Lennard Jones' interaction encounter unlike the other two water models. In order to accurately depict the ice and liquid water surrounding the melting point, a six-site water model was created. Ice and water close to the melting point were properly modeled in terms of their structural and thermodynamic properties.

4.1.1.9. Molecular Dynamics steps:

To establish the molecular system involved in the four phases (**Figure 4.4**):

1. Energy Minimization
2. Heating Dynamics
3. Equilibration Dynamics
4. Production Dynamics

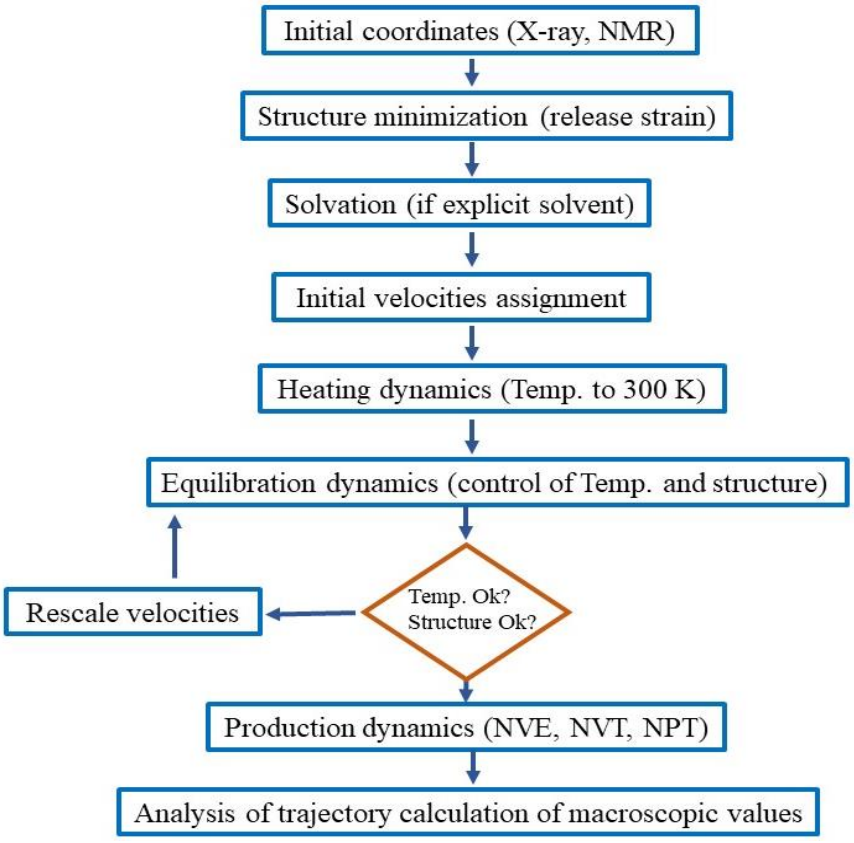


Figure 4.4. *Schematic representation of the steps involved in MD simulation.*

1. Energy Minimization

Energy minimization is used to replicate the data from MC or MD simulations. Despite being crucial for determining entropy and thermodynamic averages, MC structures and dynamical ensembles are too numerous to thoroughly examine at the microscopic level. The reduced structures are an important and helpful place to start for structural study, even though they depict the underlying configurations related to the fluctuations that happen during dynamics [25, 26]. Locating a stable point or minimum on the potential energy surface using the force field assigned to the system's atoms is required to start dynamics. The net force acting on each atom disappears on the surface of the least potential energy. Limitations can be applied in the dynamics as well as minimization processes. These restrictions can be imposed by a template to force a ligand to find the structure that is structurally closest to a target molecule, or they can be obtained from data, such as NOEs from an NMR experiment. A function (provided by the force field) and an initial estimate or set of coordinates are needed for minimization. The direction and size of a step (i.e., change in coordinates) required to approach a minimal configuration may be determined using the magnitude of the first derivative. Convergence can be rigorously described by its magnitude in addition to the size of the first derivative. There are two steps involved in the energy reduction of a molecule structure. The first stage is to create and evaluate an equation for a given conformation that describes the energy of the system as a function of its coordinates.

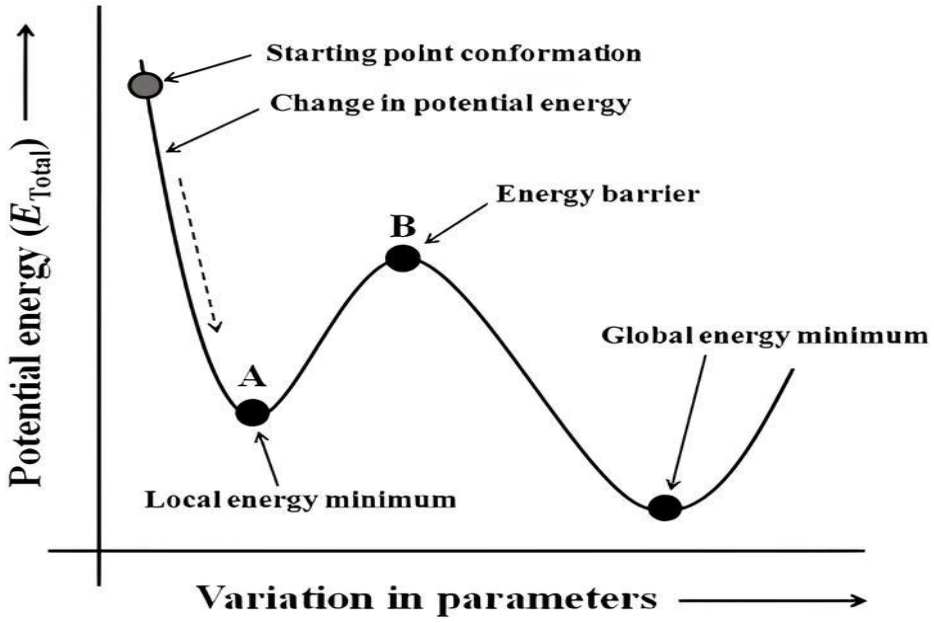


Figure 4.5. *A schematic representation of the Different phases of a molecule during the minimization of its energy (Adapted from [27])*

Figure 4.5 illustrates the several stages that an energy-saving operation might go through. Our goal is to identify the bioactive conformer since we are curious about the behavior of bioactive compounds. Research has indicated that the bioactive conformer may not be the same as the most active conformer despite the latter appearing to be physiologically potent. The bioactive conformer, however, continues to be in a zone in proximity to the most active conformer. The conformation of a ligand attached to a receptor pocket, as established by experimental studies such as X-ray crystallographic analysis, is generally regarded as the bioactive molecular arrangement or conformation if the cocrystal geometry of the molecule is present. The bioactive conformer can be regarded as the most stable conformer when there is no cocrystal geometric structure.

However, the following formulation of the energy minimization problem might be used. It must be demonstrated that, given a function *f* and one or more independent variables (x_1, x_2..... x_i), the values of each independent variable can be found by taking the minimum value of f. For any variable, the first derivative of the function at its lowest point is 0, while its second derivatives are positive:

$$\frac{\partial \mathrm{f}}{\partial x_i} = 0; \frac{\partial^2 \mathrm{f}}{\partial x_i^2} > 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.21})$$

The direction for the energy of the initial derivative determines the position of the minimum, while the gradient's magnitude informs the local slope is steep. By allowing each atom to move in response to the force applied to it, it is feasible to lower the system's energy when the force equals minus the gradient. Together with information that may be used to predict when the function will change direction (by passing through a minimum or another stationary point), the second derivatives provide information about the curvature of the function. The two techniques most commonly used in molecular modeling for the first-order minimization processes are the steepest descents and conjugate gradient approaches.

When the derivatives are near zero, minimum energy converges. It is essential to carry out energy reduction on the structure before starting an MD simulation to get rid of bad connections that might otherwise cause structural deformation. The three main minimization techniques are Newton-Raphson, conjugate gradient, and steepest descent.

(i) The Steepest Descents Method:

This approach determines which path leads to the minimum by taking the first derivative. It travels in a direction parallel to the net force. This direction is represented as a 3N-dimensional unit vector with 3N Cartesian coordinates, namely after determining the direction of travel, the next step is to determine the distance to be travelled along the gradient. In **Figure 4.6**, the two-dimensional energy surface. The gradient's direction from the beginning is along the line. The function will pass through a minimum and then rise if we visualize slicing through the surface along the line.

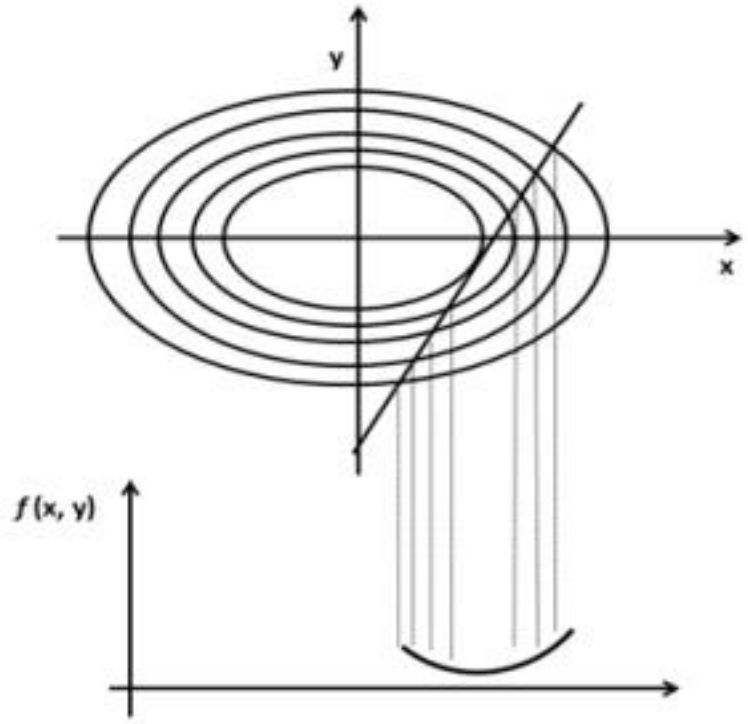


Figure 4.6. *A line search is used to locate the minimum in the function in the direction of the gradient [28].*

(ii) Minimization of Conjugate Gradients:

In the energy minimization methodology, the conjugate gradient approach yields a set of directions that do not display the steepest descents or oscillate behaviour in the narrow valleys. In the conjugate gradient approach, the gradients are orthogonal at each point, even if the directions are conjugate. Given a quadratic function with M variables, the minimum will be found in M steps from a collection of conjugate directions. Beginning at point $\boldsymbol{x}_k$, the conjugate gradient approach proceeds in the direction v_k. The gradient at the point and the preceding direction vector v_k-1 are used to determine v_k [29].

$$V_k = g_k + \gamma_k v_{k-1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.22})$$

The scalar constant γ_k in the above **equation 4.22** is provided by:

$$\gamma_k = \frac{g_k \cdot g_k}{g_{k-1} \cdot g_{k-1}} \ldots\ldots\ldots\ldots\ldots (\mathbf{4.23})$$

(iii) The Newton-Raphson method:

In the Newton-Raphson method, both the first and second derivatives are used. The curvature is utilized not just to use gradient information but also to predict where the direction will change along the function's gradient. In order to achieve energy reduction, this technique requires the greatest amount of computing power. If more water molecules are required to increase the system's solubility prior to minimization, they are introduced. An expansive container of water that has been preheated to an identical temperature is utilized for solvation. All of the system is covered by the water box, which removes any water molecules that come into contact with proteins. The mathematical model of Newton-Raphson **equation 4.24** is as follows:

$$r_{min} = r_0 - A_0^{-1.\nabla V(r_0)} \ldots\ldots\ldots\ldots\ldots (\mathbf{4.24})$$

Where A_o is the matrix of second partial derivatives of the energy with respect to the coordinates at r_o (also known as the Hessian matrix), r_{min} is the anticipated minimum, r_o is an arbitrary starting point, and $\nabla V(r_o)$ is the gradient of the potential energy at r_o.

2. Heating Dynamics

During the heating phase, when each atom in the system is allocated a beginning velocity (at 0 K), the Newton's equations of motion, which show the temporal growth of the system, are numerically integrated. Following the assignment of new velocities at short, predefined intervals that correspond to marginally higher temperatures, the simulation is set to continue until the goal temperature of 300 K is attained. Heating causes structural stresses to relax, which in turn releases force limitations on different subdomains of the simulation system. Constant volume (NVT) is the typical working condition for thermal dynamics.

3. Equilibration Dynamics

During the equilibrium phase, a system reaches equilibrium as it changes from its starting state. Equilibration should continue continuously, or at least until the values of the set of monitored attributes are settled. Together with structural features, the main measured attributes are thermodynamic variables like energy, temperature, and pressure. Nonetheless, the initial structure of the liquid state simulations is similar to a solid lattice. In actuality, it is imperative to arrange the components so that the lattice has "melted" prior to the start of the manufacturing process. The reaching point of the liquid state can be ascertained by using the order parameters. This order parameter refers to the assessment of a system's degree of order. The atoms might, however, remain mostly in the same location throughout, preserving a high degree of order while mimicking a crystal lattice. Translational dysfunction may result from the species' propensity for frequent movement inside a liquid. Solving the equations of motion for an atomic system is known as MD. The equation of molecular motion can be solved to determine its trajectory and the temporal development of its motions. MD allows for the bridging of barriers and the examination of several alternatives, depending on the temperature at which the simulation is run. First, velocities need to be assigned in order to start the MD. The Maxwell-Boltzmann distribution limitation for the random number generator is used to achieve this. The average kinetic energy of the system establishes the temperature, according to the kinetic theory of gases. The internal energy of the system is measured in $U = 3/2\ NkT$.

$U = 1/2\ Nmv^2$ gives the kinetic energy of the system. However, the temperature may be determined by taking the average of all the atoms' velocities in the system. Throughout the simulation, the Maxwell-Boltzmann distribution may be maintained after the starting set of velocities is determined. The temperature can be considered to be strictly zero Kelvin after energy reduction. Before the system can be heated to the required temperature, the dynamics must be initialized. In order to assign velocities at low temperatures, dynamics is performed in compliance with the equations of motion. The temperature is raised after a number of dynamics rounds, though. A 20 ps timeline of progressive heating dynamics is conducted from 0 to 300 K under atomic constraints. Velocity scaling is the most often used approach to temperature scaling. A run of at least 5 ps (5000 time steps) and frequently 10 or 20 ps is necessary for equilibration time steps of 1 fs. After heating, dynamic equilibration lasts for 100 ps.

4. Production Dynamics:

For the dynamics of the time period of interest is mostly employed to compute thermodynamic averages or sample new configurations. During this step, calculations are performed about the thermodynamic properties and further data. The simulated system determines most of the parameters, including the kinetic, potential, and total energy, velocities, temperature, and pressure, which are used to determine whether or not equilibrium has been attained. However, in a simulation of the micro canonical ensemble, the total energy remains constant, even though the kinetic and potential energies could change. Each of the three directions, x, y, and z, should possess an equivalent quantity of kinetic energy, and the velocity components need to fall inside the Maxwell-Boltzmann distribution. The system's variable during the production phase is the temperature. The system is left to evolve when all counts are reset to zero at the beginning of the manufacturing phase. The temperature of the system is now estimated because no velocity scaling is carried out while creating a micro canonical ensemble. The characteristics are precisely calculated and retained for additional processing and analysis throughout the manufacturing phase. If problems occur, it could be required to restart the simulation if it’s being closely observed based on its behavior. Additionally, it is standard procedure to save configurations' energies, positions, and velocities throughout time in order to retrieve the other properties when the simulation is finished. Nevertheless, the MD simulation may be performed while the thermodynamic parameters are being calculated. The production run is created using a few hundred ps to ns or even more.

4.1.2. Molecular docking:

The computational anticipation of interactions between protein molecules poses a formidable task within the domain of structural biology. Accurate and dependable prediction of these interactions holds substantial promise for advancing various realms of biological research, spanning academia and industry alike. The complexity inherent in protein-protein docking lies in the precise alignment of two interacting molecules, contingent upon the interactions among constituent residues engaged in the targeted interaction. Multiple docking methodologies are currently accessible to address this challenge [30-33].

4.1.2.1. ClusPro web server:

In 2004, the ClusPro web-based server was initially introduced [34, 35], subsequently undergoing notable refinement and expansion[36-38]. ClusPro facilitates the direct docking of two interacting proteins [39], requiring two protein data bank (PDB)-formatted files for execution. The server employs a three-step computational process for docking:

(i) Rigid body docking is employed to explore billions of conformations.

(ii) Subsequently, 1000 lowest energy structures are clustered based on RMSD to identify prominent clusters reflecting plausible complex models;

(iii) Energy minimization is employed to refine selected structures (**Figure 4.7**). During the rigid body docking phase, PIPER, a docking tool, utilizes the FFT correlation technique [40].

Presently, the ClusPro web server has transitioned to ClusPro 2.0, reflecting an updated version.

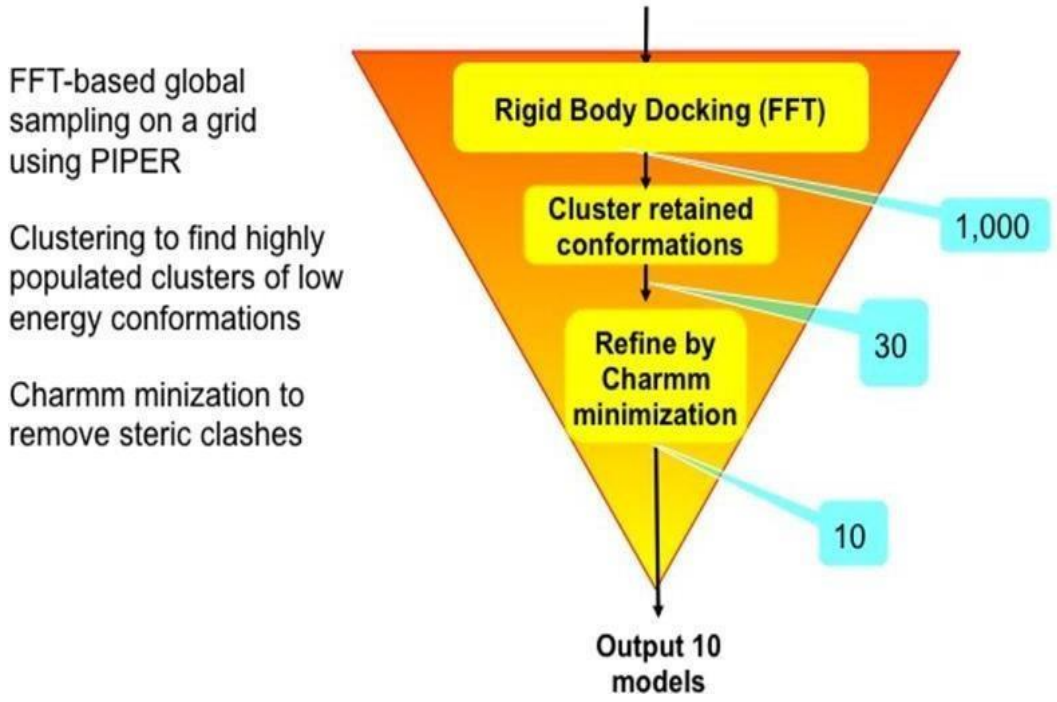


Figure 4.7. *Representation of the ClusPro algorithm: the number of structures retainedafter each step is shown in a blue box. Taken from [39].*

4.1.3. In silico prediction of protein-protein interaction:

Protein-protein interactions (PPIs) are an important mechanism that drives a variety of cellular physiological functions and are also implicated in the pathophysiology of many illnesses [41, 42].

4.1.3.1. PDBsum server:

A web-based database called PDBsum provides a visual summary of all the important information pertaining to every macromolecular structure that has been submitted to the Protein Data Bank (PDB) (http://www.ebi.ac.uk/pdbsum) [43]. Included are extensive structural pictures, annotated plots of each protein chain's secondary structure, thorough structural analyses, a synopsis of the PROCHECK findings, and schematic diagrams of the interactions between proteins and small molecules, proteins and DNA, and other molecules. Important structural elements like the protein's domains, PROSITE patterns, and interactions between proteins and ligands are highlighted using RasMol scripts. Publicly accessible at http://www.biochem.ucl.ac.uk/bsm/pdbsum, PDBsum is updated whenever new structures are made available by the PDB. This server produced a PDB structural information library, making it the first webserver to utilize the new World Wide Web technology. Its main goal was to offer a comprehensive visual encyclopedia of the proteins and their complexes included in the PDB. It was first developed in 1995 at University College London (UCL). These images are made up of many structural studies that are either not available or not easily accessible elsewhere.

Providing each 3-D model's structural information in the most visually appealing way possible is the main goal of this server. The molecules that make up each PDB entry, such as ligands, metals, protein/DNA/RNA chains, and their interactions, are thus shown graphically. Over time, a growing number of new features have been introduced. One may painstakingly compile this kind of material for oneself, in addition to the references to literature and other links to databases; however, it's preferable to deliver it right away. The following features of the PDBsum server include:

a) **Wiring Diagram**

PDBsum offers a "protein page" with a schematic representation of the protein's secondary structure, or "wiring diagram," for each distinct protein chain of a structural model (**Figure 4.8**).

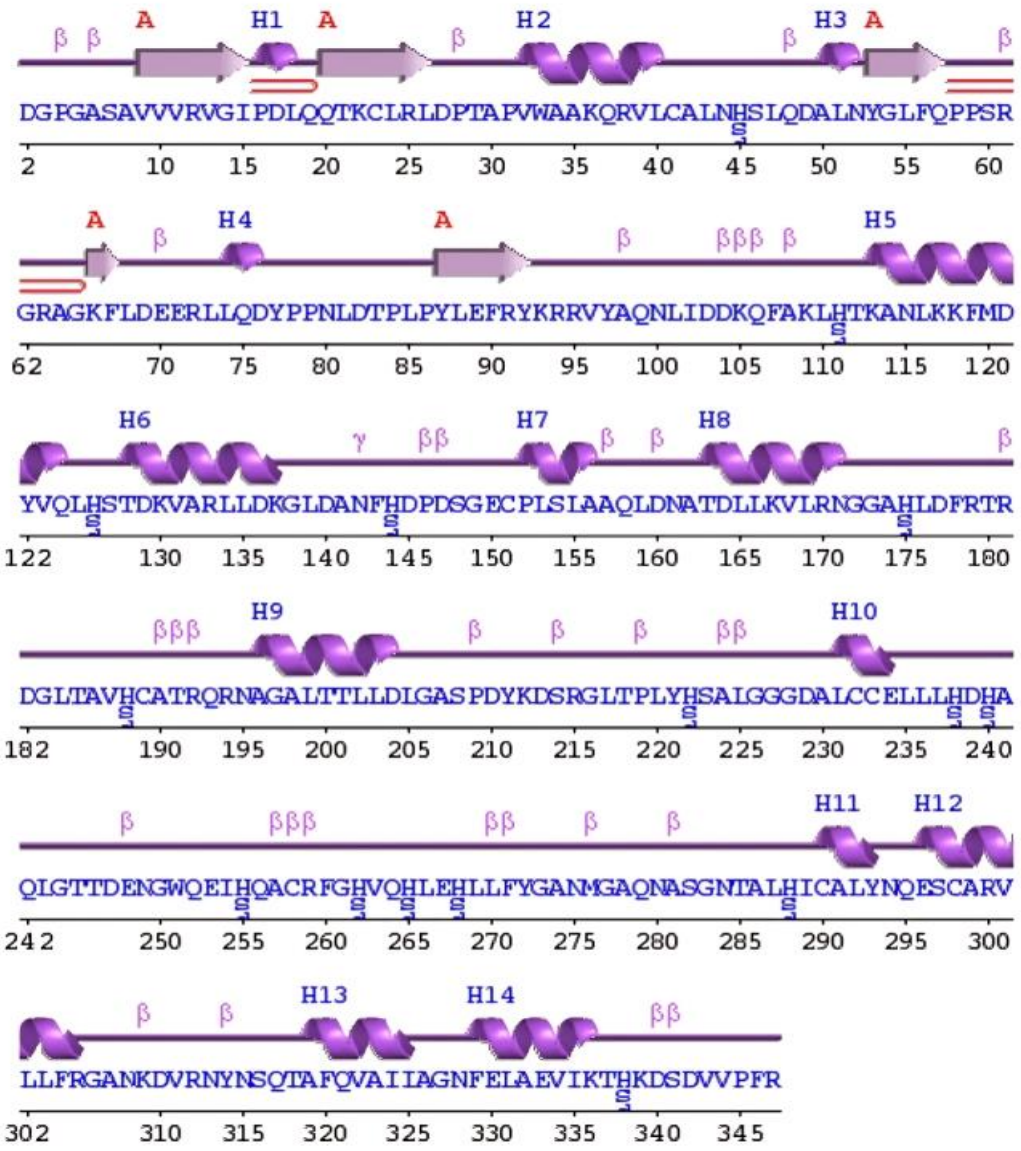


***Figure 4.8**. Wiring diagram presents the secondary structures in the SHANK3 protein (5G4X pdb). Helices are labelled as H1, H2, H3, H4, H5, H6, H7, H8, H9, H10, H11, H12, H13 and H14.*

b) Surface topology

The protein page also includes a topology diagram (**Figure 4.9**) that illustrates the connections and arrangements between the protein's strands and helices. When a protein chain includes many domains, the wiring diagram's domain shading is followed in the creation of each domain's diagram, which is done independently. Hydrogen bonding plots may be converted into topology diagrams using Gail Hutchinson's HERA programme [44].

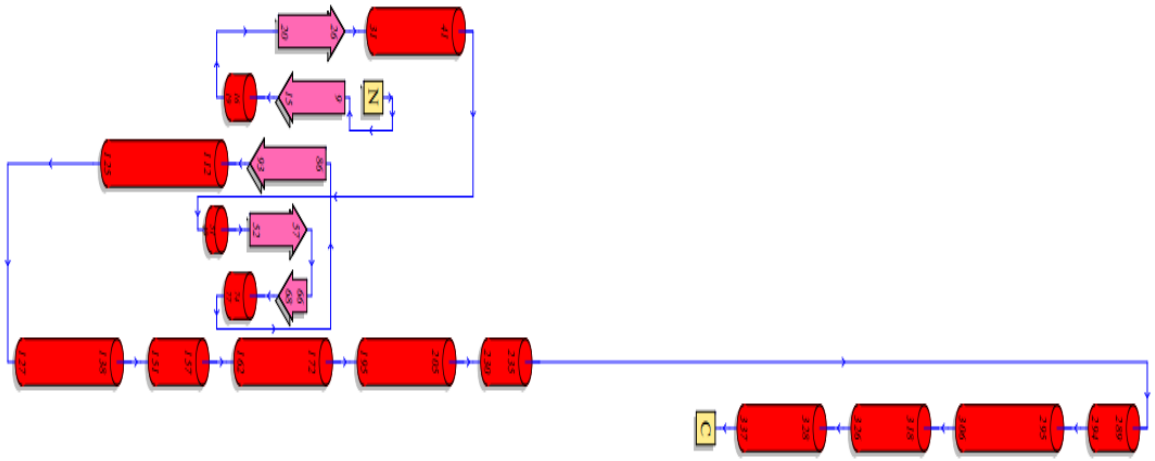


Figure 4.9. *The surface topology diagram delineates the distribution of structural motifs: helices, beta turns, and gamma turns in SHANK3 protein (5G4X pdb).*

4.1.4. YASARA tool:

YASARA tool [45, 46] is an application designed for protein molecular modeling that is needed for this kind of computational work. Molecular or homology-based modeling are two computer approaches that can produce structural information about the reaction of the system. Two particular applications are demonstrated, encompassing both homology modeling and molecular modeling techniques like energy minimization, molecular docking simulations, and MD simulations. The applications have been selected to provide concrete illustrations of how structural information obtained from homology and molecular modeling is applied to direct protein modeling research studies.

4.1.5. Analysis of MD trajectory:

(i) Root Mean Square Deviation (RMSD):

RMSD is a measurement used to determine a structure's deviation from a certain conformation. It is described as:

$$RMSD = \left[\frac{\sum_N \left(R_i - {R_i}^0\right)^2}{N}\right]^{1/2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.25})$$

Where R_i is the vector location of particle i (the target atom) in the snapshot, R_i^0 is the coordinate vector for reference atom i, and N is the total number of atoms/residues taken

into account in the computation. Using backbone atoms and the simulation's first frame as a reference, the RMSD was calculated. The RMSD is the product of the number of locations (i), the number of strands (j), and the number of angular parameters (k). The value of N in **equation 4.25** denotes the total number of variables needed to compute the RMSD. A radial vector of length r in the structure space denoted by the RMSD absolute magnitude is the calculated RMSD. The radial issue is predicated on the idea that there is more configurational space volume between a given r and r+dr, the broader the radius. The same RMSD value might capture both equivalent and different structures at larger r values. Techniques that use significantly lower RMSD values offer a more precise way to quantify differences. The existence of two or more structural substrates creates a second important problem with the use of RMSD because of the molecule's intrinsic flexibility. However, a technique is required to define the dynamic properties accurately without compromising the information.

(ii) Root Mean Square Fluctuation (RMSF):

The measure of divergence between the particle location i and a reference position is defined by Root Mean Square Fluctuation or RMSF:

$$\mathrm{RMSF} = \left[\frac{1}{\mathrm{T}}\sum_{\mathrm{t=1}}^{\mathrm{T}}(r_i(t) - r_i^{ref})^2\right]^{1/2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.26})$$

T is the desired average time, and r_i^{ref} is the particle i reference location, as shown in **equation 4.26**. The time-averaged location will serve as the reference location of the same particle i, that is, $r_i^{ref} = r_i$. Root mean square deviation (RMSD) and root mean square fluctuation (RMSF) are metrics used to quantify the spatial fluctuations of biomolecules in MD simulations. For a given collection of atoms, RMSD is the difference between two structures; on the other hand, RMSF is the variation around an average, per atom or residue, over a series of structures (e.g., from a trajectory).

(iii) Radius of gyration (Rg):

The radius of gyration is computed to determine the structure's compactness:

$$\mathrm{R_g} = \left[\frac{\sum_i |\mathrm{r_i}| 2\mathrm{m_i}}{\sum_i \mathrm{m_i}}\right]^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.27})$$

The mass of atom i is denoted by m_i in **equation 4.27** and its location in relation to the molecule's centre of mass is indicated by r_i.

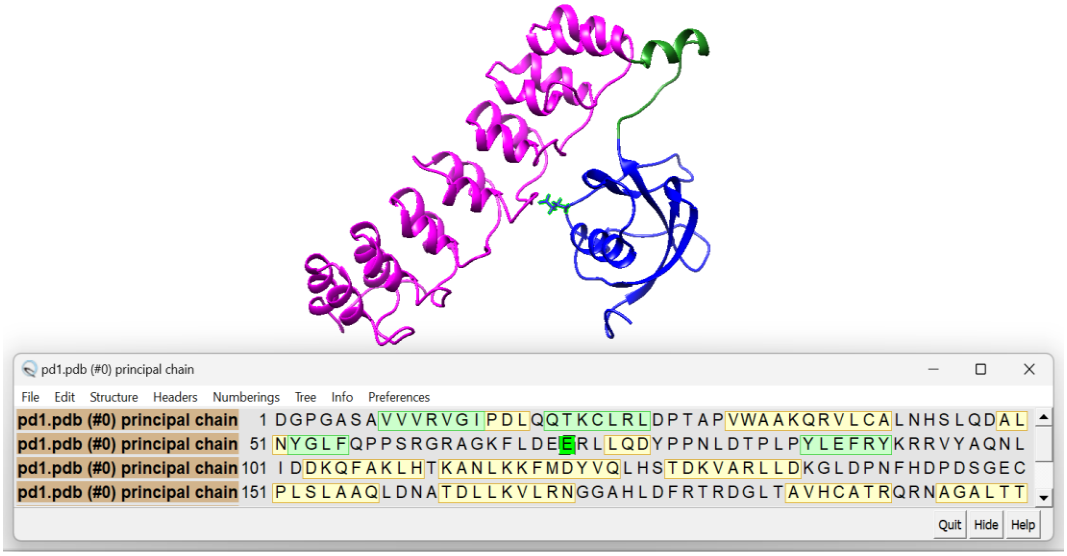
(iv) DSSP plot analysis : Secondary structural content analysis:

For most of the proteins in the PDB, Kabsch and Sander (1983) created a database of the ASA by using a method called Dictionary of Secondary Structure for Protein (DSSP) to ascertain the solvent accessibility of the residues [47]. This programme is frequently used to create the ASA values for use in prediction algorithms [288-290]. It works by primarily classifying secondary structures of proteins based on their backbone H-bonds. It also gives information on bond and Cα-pseudo dihedral angles; only the latter is required by DSSP in order to determine the LSS of a residue. The electrostatic hydrogen bond detection criteria set it apart. Consequently, the unique hydrogen-bond patterns are being used to designate the elements of the secondary structure. This method is often used to assign secondary structures as a gauge. DSSP is used by many software applications to assign secondary structures as needed. For example, a widely used visualization tool such as Rasmol assigns repeating structures in a fast way that is comparable to the DSSP. Classification of β-bridges is based on non-recurring H-bonds, whereas classification of helix or strands is based on repeating patterns of the same type of H-bonds [48]. Residues that share the same secondary structure pattern are clustered together quite tightly in a Ramachandran plot because the relative orientations of the nitrogen and oxygen atoms in the backbone are reflected in the corresponding (φ, ψ) backbone tilt angles. The DSSP offers information on the secondary structure of the protein on two levels [47]. The one-character secondary structure information (1CSSI) code, which at the higher level characterizes the LSS of a residue mostly based on the H-bond arrangement of the protein backbone into eight classes, summarises the secondary structure from DSSP analysis. The angle that exists between the vectors Cα (i) − Cα (i−2) and Cα (i) − Cα (i+2) for each residue is known as the C-pseudo bond angle, and it is instead calculated using DSSP. A gap for better discrimination is created by DSSP in the event that none of the previously specified conditions are satisfied; this gap is represented by the letter "C." These residues, however, correspond to a usually straight section of the protein backbone structure and do not contain the backbone H-bonds required for the formation of secondary structures.

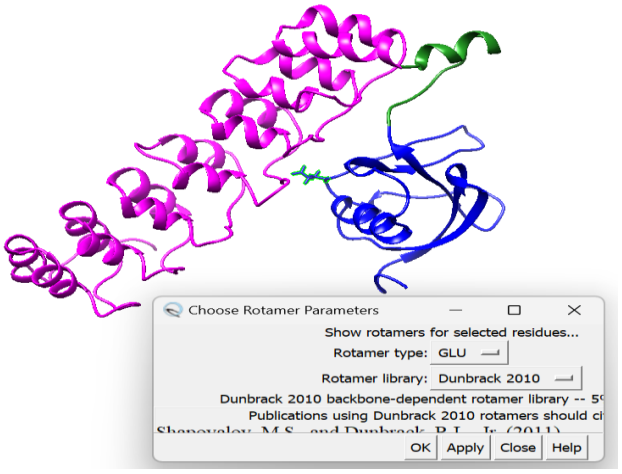
4.1.6. 3-D structure visualization tool:

4.1.6.1. UCSF Chimera:

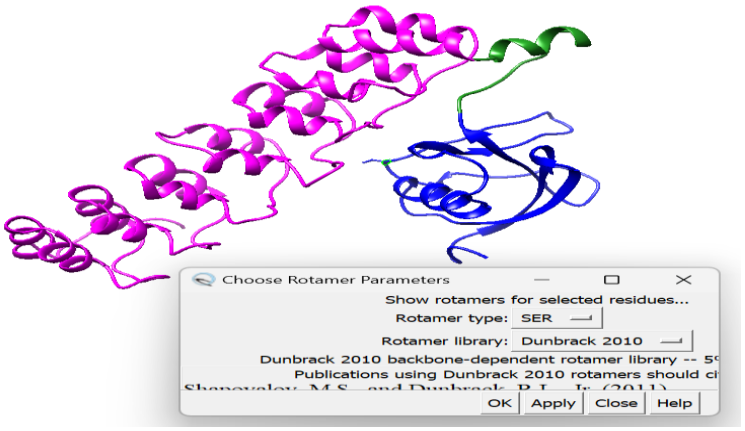
University of California, San Francisco (UCSF) Chimera: A highly adaptable instrument known as UCSF Molecular structures and related data, including conformational ensembles, density maps, sequence alignments, supramolecular assemblies, and docking results, may interactively seen and analyzed using Chimera [49]. Chimera is an application that was created by the Resource for Biocomputing, Visualisation, and Informatics (RBVI) with support from the National Institutes of Health (NIH). (**Figure 4.10**) present the steps of construction SHANK3 mutants utilizing Chimera.



STEP 1: Tools → Sequence → Sequence → Select the residue (Cltr + shift + left click)



STEP 2: Tools → Structure Editing → Rotamers → Rotamer type → Choose the mutant residue



STEP 3: Mutated residue → Apply → Ok

Figure 4.10. *Construction of SHANK3 mutants using UCSF Chimera software [49].*

4.1.7. R programming:

The R language came into use quite a bit after S had been developed. One key limitation of the S language was that it was only available in a commercial package, S-PLUS. In 1991, R was created by Ross Ihaka and Robert Gentleman in the Department of Statistics at the University of Auckland. In 1993, the first announcement of R was made to the public [50]. Later, R software became free, and it is critical because it allowed the source code for the entire R system to be accessible to anyone who wanted to tinker with it. In 2000, version 1.0.0 of R was released to the public.

4.1.7.1. DESeq2 package to detect differentially expressed genes (DEGs):

A fundamental aspect of RNA-seq data analysis involves discerning variances in gene expression levels. Typically, count data are organized in tabular form, detailing the quantity of sequence fragments attributed to individual genes across different samples. A pivotal analytical inquiry pertains to quantifying systematic alterations between experimental conditions while considering variability inherent within each condition. DESeq2, a software package, offers methodologies for examining differential expression through negative binomial generalized linear models. These models leverage data-driven prior distributions to estimate dispersion and logarithmic fold changes, facilitating robust statistical inference.

4.1.7.1.1. Theory behind DESeq2 model

The DESeq2 model for differential expression analysis utilizes a generalized linear model of the form [51]:

$$K_{ij} \sim \text{NB}\left(\mu_{ij}, \alpha_i\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.29})$$

$$\mu_{ij} = s_j\, q_{ij} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.30})$$

$$\log_2(q_{ij}) = x_j.\, \beta_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{4.31})$$

where raw counts K_{ij} for gene i, sample j are modeled using a negative binomial distribution with fitted mean μ_{ij} and a gene-specific dispersion parameter α_i. The fitted mean is composed of a sample-specific size factor s_j and a parameter q_{ij} proportional to the expected true concentration of fragments for sample j [52]. The coefficients β_i give the $\log_2$ fold changes for gene i for each column of the model matrix X. The dispersion parameter α_i elucidates the association between the variance exhibited by the observed count and its corresponding mean value. Essentially, it quantifies the degree of deviation anticipated for the observed count relative to its mean, which is based both on the size factor s_j and the covariate-dependent part q_{ij} as previously described.

$$\text{Var}\left(K_{ij}\right) = E\left[\left(K_{ij\,-}\,\mu_{ij}\right)^2 = \mu_{ij} + \alpha\mu_{ij}^2\right] \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (\mathbf{4.32})$$

The $\log_2$ fold changes in β_i are the maximum a posteriori estimates after incorporating a zero-centered Normal prior – in the software referred to as a β-prior – hence DESeq2 provides "moderated" $\log_2$ fold change estimates. Dispersions are estimated using expected mean values from the maximum likelihood estimate of $\log_2$ fold changes, and optimizing the Cox-Reid adjusted profile likelihood [53, 54].

4.1.7.1.2. Hypothesis testing using the Wald test

The procedure of hypothesis testing commences with the formulation of a null hypothesis for each gene. In our context, the null hypothesis posits no discernible differential expression across the two sample groups, denoted by a log-fold change (LFC) equaling zero. Importantly, this hypothesis formulation is independent of any empirical data and is purely conceptual. Subsequently, statistical testing is employed to assess the veracity of the null hypothesis based on observed data. Within the framework of DESeq2, the Wald test serves as a prominent method for hypothesis evaluation in

comparisons involving two groups. This test yields a Wald test statistic, accompanied by the computation of the probability that a test statistic as extreme as the observed value could arise by chance. Termed the p-value.

4.1.7.1.3. Log2 Fold Change:

The determination of fold change entails computing the ratio of normalized read counts observed between two distinct conditions under investigation. However, expressions of gene level changes are frequently represented as $\log_2$ fold change. The utilization of logarithmic transformation proves particularly advantageous for visualizing alterations in gene expression. Moreover, it facilitates an intuitive understanding, where a $\log_2$ fold change of 1 signifies a doubling of expression level within a specific condition. In contrast, a negative $\log_2$ fold change indicates down-regulation of the gene in said condition.

In the latest versions of DESeq2, default settings no longer include shrinkage of $\log_2$ fold change estimates. Consequently, the $\log_2$ fold changes would be the similar to those computed by normalized_counts_group1 / normalized_counts_group2, as shown in equation (4.33)

$$\log_2(normalized\_counts\_group1 \,/\, normalized\_counts\_group2) \;\cdots\cdots\cdots (4.33)$$

4.1.7.1.4. The probability value (P-value)

The P-value is the probability, which signifies the likelihood of observing the test statistic under the null hypothesis. A small p-value prompts the rejection of the null hypothesis, indicating substantive evidence against it, thereby implying differential expression of the gene in question. The conventional practice of employing a p-value threshold, often set at 0.05, to discern statistically significant findings is commonplace when scrutinizing individual genes. Conceptually, a p-value of 0.05 signifies a 5% probability that the observed difference is attributable to chance, thereby denoting a false-positive result. However, when examining the transcriptome, wherein analyses encompass numerous genes simultaneously, the cumulative risk of encountering false-positive outcomes escalates significantly. For instance, considering a dataset comprising 10,000 genes, the potential number of false-positive hits would amount to 0.05 * 10,000 = 500 occurrences. Such an abundance of ostensibly significant findings undermines the reliability of the results, constituting what is termed a multiple-testing problem —

wherein the likelihood of obtaining positive outcomes purely by chance increases with the number of tests conducted. To mitigate this issue, an alternative metric, termed the q-value, is utilized in lieu of the p-value.

4.1.7.1.5. False Discovery Rate (FDR)

The adjusted p-value, commonly referred to as the q-value or FDR value, is a metric derived from the p-value. It is computed to address the issue of multiple testing, wherein a large number of statistical tests are conducted simultaneously. The FDR value quantifies the proportion of false positives among the results deemed statistically significant, considering the entire set of significant findings. A lower FDR value indicates greater confidence in the significance of the observed results, as it reflects a lower likelihood of false positives among the identified significant findings. DESeq2 calculates the FDR value using methods such as the Benjamini-Hochberg procedure to adjust the p-values obtained from hypothesis testing, thereby controlling the rate of false positives while identifying differentially expressed genes. Typically, the 0.05 value warrants attention when assessing the presence of differential gene expression.

4.1.7.2 Enrichment Functional Analysis

4.1.7.2.1. ClusterProfiler package

Pathway enrichment analysis constitutes a pivotal endeavor in discerning the underlying biological themes inherent in high-throughput sequencing data. The clusterProfiler package provides a variety of functions that reveal biological processes and trajectories.

[55]. In 2012, the clusterProfiler package was initially released [56]. Primarily designed to execute over-representation analysis (ORA) [57] utilizing Gene Ontology (GO) [58] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [59] annotations across diverse model organisms, ClusterProfiler facilitates the comparative examination of functional profiles under varying experimental conditions [53]. Over time, clusterProfiler has undergone substantial refinement, extending its support to encompass diverse ontology and pathway annotations, along with updated gene annotations for thousands of species. Additionally, users have the capability to incorporate custom annotation data for novel species alongside accommodating emerging annotations. Notably, clusterProfiler accommodates both ORA and gene set enrichment analysis (GSEA) [60] are supported. A complicated experimental design that enables comparison of functional profiles of

diverse circumstances on different levels is supported by an extension of the comparative utility.

(i) Gene Ontology

The Gene Ontology (GO) categories comprise precisely delineated terms that encapsulate various properties of gene products. This ontology is structured across three fundamental domains, as elucidated by Ashburner et al. (2000) [58]:

- Cellular component: Encompasses the constituents of a cell or its extracellular environment.
- Molecular function refers to a gene product's basic molecular functions, including interaction and catalysis.
- Biological process: Signifies operations or collections of molecular events characterized by distinct initiation and termination points integral to the operational dynamics of integrated living entities, such as cells, tissues, organs, and organisms.

(ii) Kyoto Encyclopedia of Genes and Genomes (KEGG)

Kyoto Encyclopedia of Genes and Genomes serves as a comprehensive repository of genetic and genomic information [59]. Within KEGG, molecular functions are depicted through networks of interactions and reactions primarily structured as KEGG pathways and modules. A KEGG module is a compilation of functionally defined units, offering a coherent representation of biological processes. Notably, both KEGG pathways and modules find support within the clusterProfiler package. However, it is noteworthy that several software tools supporting KEGG analysis ceased updates following KEGG's transition to an academic subscription model in July 2011. Consequently, these tools may employ outdated KEGG data, potentially yielding inaccurate or misleading results.

Fortunately, the KEGG web resource remains freely accessible, mitigating concerns associated with outdated data. Notably, the clusterProfiler package circumvents the need to package KEGG data internally. Instead, it dynamically queries the latest KEGG database online through web API to conduct functional analysis. This approach offers a distinct advantage, enabling clusterProfiler to utilize current data while accommodating all species endowed with KEGG annotation, exceeding 6,000 species [55].

4.1.7.3. The Database for Annotation, Visualization and Integrated Discovery (DAVID)

DAVID, a freely accessible online bioinformatics resource, has been developed by the Laboratory of Human Retrovirology and Immunoinformatics (LHRI) [61, 62]; the suite of tools within the DAVID Bioinformatics Resources is specifically designed to facilitate the functional interpretation of extensive gene lists stemming from genomic investigations, such as microarray and proteomics studies. DAVID's functionalities are accessible at https://david.ncifcrf.gov/.

DAVID provides a comprehensive set of functional annotation tools for investigators to understand the biological meaning behind a large list of genes. For any given gene list, DAVID tools are able to:

- Identify enriched biological themes, particularly GO terms
- Uncover enriched gene groups associated with specific functional themes. Cluster redundant annotation terms
- Visualize gene associations within KEGG pathway maps.
- Explore additional functionally related genes not present in the initial list.
- List interacting proteins
- Link gene-disease associations
- Convert gene identifiers from one type to another.

DAVID offers a comprehensive array of functional annotation tools tailored to assist researchers in elucidating the biological significance underlying large gene lists. Specifically, DAVID's tools enable investigators to [59]:

References

[1] Du, X., Li, Y., Xia, Y.-L., Ai, S.-M., Liang, J., Sang, P., Ji, X.-L. and Liu, S.-Q. Insights into protein–ligand interactions: mechanisms, models, and methods. *International journal of molecular sciences,* 17: 144-178, 2016. https://doi.org/10.3390/ijms17020144

[2] Hollingsworth, S. A. and Dror, R. O. Molecular dynamics simulation for all. *Neuron,* 99: 1129-1143, 2018. https://doi.org/10.1016/j.neuron.2018.08.011

[3] McCammon, J. A., Gelin, B. R. and Karplus, M. Dynamics of folded proteins. *nature,* 267: 585-590, 1977. https://doi.org/10.1038/267585a0

[4] Levitt, M. and Huber, R. Molecular dynamics of native protein: II. Analysis and nature of motion. *Journal of molecular biology,* 168: 621-657, 1983. https://doi.org/10.1016/S0022-2836(83)80306-4

[5] Karplus, M. and McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nature structural biology,* 9: 646-652, 2002. https://doi.org/10.1038/nsb0902-646

[6] Alder, B. J. and Wainwright, T. E. Phase transition for a hard sphere system. *The Journal of chemical physics,* 27: 1208, 1957. https://doi.org/10.1063/1.1743957

[7] Alder, B. J. and Wainwright, T. E. Studies in molecular dynamics. I. General method. *The Journal of Chemical Physics,* 31: 459-466, 1959. https://doi.org/10.1063/1.1730376

[8] Rahman, A. Correlations in the motion of atoms in liquid argon. *Physical review,* 136: A405, 1964.

[9] Rahman, A. and Stillinger, F. H. Molecular dynamics study of liquid water. *The Journal of Chemical Physics,* 55: 3336-3359, 1971. https://doi.org/10.1063/1.1676585

[10] McCAMMON, J. A., Karim, O. A., Lybrand, T. P. and Wong, C. F. Ionic Association in Water: From Atoms to Enzymes a. *Annals of the New York Academy of Sciences,* 482: 210-221, 1986. https://doi.org/10.1111/j.1749-6632.1986.tb20952.x

[11] Duan, Y. and Kollman, P. A. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science,* 282: 740-744, 1998. https://DOI.org/10.1126/science.282.5389.74

[12] Heidari, Z., Roe, D. R., Galindo-Murillo, R., Ghasemi, J. B. and Cheatham III, T. E. Using wavelet analysis to assist in identification of significant events in molecular dynamics simulations. *Journal of chemical information modeling,* 56: 1282-1291, 2016. https://doi.org/10.1021/acs.jcim.5b00727

[13] Adcock, S. A. and McCammon, J. A. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chemical reviews,* 106: 1589-1615, 2006. https://doi.org/10.1021/cr040426m

[14] Chang, C.-E. A., Huang, Y.-M. M., Mueller, L. J. and You, W. Investigation of structural dynamics of enzymes and protonation states of substrates using computational tools. *Catalysts,* 6: 82, 2016. https://doi.org/10.3390/catal6060082

[15] Hummer, G. The numerical accuracy of truncated Ewald sums for periodic systems with long-range Coulomb interactions. *Chemical physics letters,* 235: 297-302, 1995. https://doi.org/10.1016/0009-2614(95)00117-M

[16] Ryckaert, J.-P., Ciccotti, G. and Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of computational physics,* 23: 327-341, 1977. https://doi.org/10.1016/0021-9991(77)90098-5

[17] Case, D. A., Darden, T. A., Cheatham, T. E., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Crowley, M., Walker, R. C. and Zhang, W. Amber 10. 2008.

[18] Bernal, J. D. and Fowler, R. H. A theory of water and ionic solution, with particular reference to hydrogen and hydroxyl ions. *The Journal of Chemical Physics*, 1: 515-548, 1933. http://dx.doi.org/10.1063/1.1749327

[19] Guillot, B. A reappraisal of what we have learnt during three decades of computer simulations on water. *Journal of molecular liquids,* 101: 219-260, 2002. https://doi.org/10.1016/S0167-7322(02)00094-6

[20] Berendsen, H.-J.-C., Grigera, J.-R. and Straatsma, T. P. The missing term in effective pair potentials. *Journal of Physical Chemistry,* 91: 6269-6271, 1987. https://doi.org/10.1021/j100308a038

[21] Mahoney, M. W. and Jorgensen, W. L. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *The Journal of chemical physics,* 112: 8910-8922, 2000. https://doi.org/10.1063/1.481505

[22] Horn, H. W., Swope, W. C., Pitera, J. W., Madura, J. D., Dick, T. J., Hura, G. L. and Head-Gordon, T. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *The Journal of chemical physics,* 120: 9665-9678, 2004. https://doi.org/10.1063/1.1683075

[23] Horn, H. W., Swope, W. C. and Pitera, J. W. Characterization of the TIP4P-Ew water model: Vapor pressure and boiling point. *The Journal of chemical physics,* 123, 2005. https://doi.org/10.1063/1.2085031

[24] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. and Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics,* 79: 926-935, 1983. https://doi.org/10.1063/1.445869

[25] Hagler, A., Theoretical simulation of conformation, energetics, and dynamics of peptides. In *Conformation in Biology and Drug Design*, Elsevier: 1985; pp 213-299.

[26] Struthers, R., Hagler, A. and Rivier, J., Design of peptide analogs: Theoretical simulation of conformation, energetics, and dynamics. ACS Publications: 1984; pp 239-261.

[27] Seifoori, S., Ebrahimi, F., Parrany, A. M. and Liaghat, G. Dynamic analysis of single–layered graphene sheet subjected to a moving nanoparticle: A molecular dynamics study. *Materials Science Engineering: B,* 285: 115956, 2022. https://doi.org/10.1016/j.mseb.2022.115956

[28] Wright, S. J., Numerical optimization. 2006; pp 1-653.

[29] Kirkwood, J. G. Statistical mechanics of fluid mixtures. *The Journal of chemical physics,* 3: 300-313, 1935. https://doi.org/10.1063/1.1749657

[30] Camacho, C. J., Gatchell, D. W., Kimura, S. R. and Vajda, S. Scoring docked conformations generated by rigid-body protein-protein docking. *Proteins: Structure, Function, Bioinformatics,* 40: 525-537, 2000. https://doi.org/10.1002/1097-0134(20000815)40:3<525::AID-PROT190>3.0.CO;2-F

[31] Goldman, B. B. and Wipke, W. T. QSD quadratic shape descriptors. 2. Molecular docking using quadratic shape descriptors (QSDock). *Proteins: Structure, Function, Bioinformatics,* 38: 79-94, 2000. https://doi.org/10.1002/(SICI)1097-0134(20000101)38:1<79::AID-PROT9>3.0.CO;2-U

[32] Gardiner, E. J., Willett, P. and Artymiuk, P. Protein docking using a genetic algorithm. *Proteins: Structure, Function, Bioinformatics,* 44: 44-56, 2001. https://doi.org/10.1002/prot.1070

[33] Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A. and Baker, D. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of molecular biology,* 331: 281-299, 2003. https://doi.org/10.1016/S0022-2836(03)00670-3

[34] Comeau, S. R., Gatchell, D. W., Vajda, S. and Camacho, C. J. ClusPro: a fully automated algorithm for protein–protein docking. *Nucleic acids research,* 32: W96-W99, 2004. https://doi.org/10.1093/nar/gkh354

[35] Comeau, S. R., Gatchell, D. W., Vajda, S. and Camacho, C. J. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics,* 20: 45-50, 2004. https://doi.org/10.1093/bioinformatics/btg371

[36] Comeau, S. R., Kozakov, D., Brenke, R., Shen, Y., Beglov, D. and Vajda, S. ClusPro: performance in CAPRI rounds 6–11 and the new server. *Proteins: structure, function, bioinformatics,* 69: 781-785, 2007. https://doi.org/10.1002/prot.21795

[37] Kozakov, D., Hall, D. R., Beglov, D., Brenke, R., Comeau, S. R., Shen, Y., Li, K., Zheng, J., Vakili, P. and Paschalidis, I. C. Achieving reliability and high accuracy in

automated protein docking: ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13–19. *Proteins: Structure, Function, Bioinformatics*, 78: 3124-3130, 2010. https://doi.org/10.1002/prot.22835

[38] Kozakov, D., Beglov, D., Bohnuud, T., Mottarella, S. E., Xia, B., Hall, D. R. and Vajda, S. How good is automated protein docking? *Proteins: Structure, Function, Bioinformatics*, 81: 2159-2166, 2013. https://doi.org/10.1002/prot.24403

[39] Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., Beglov, D. and Vajda, S. The ClusPro web server for protein–protein docking. *Nature protocols*, 12: 255-278, 2017. https://doi.org/10.1038/nprot.2016.169

[40] Kozakov, D., Brenke, R., Comeau, S. R. and Vajda, S. PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins: Structure, Function, Bioinformatics*, 65: 392-406, 2006. https://doi.org/10.1002/prot.21117

[41] Cohen, F. E. and Prusiner, S. B. Pathologic conformations of prion proteins. *Annual review of biochemistry*, 67: 793-819, 1998. https://doi.org/10.1146/annurev.biochem.67.1.793

[42] Loregian, A., Marsden, H. S. and Palu, G. Protein–protein interactions as targets for antiviral chemotherapy. *Reviews in medical virology*, 12: 239-262, 2002. https://doi.org/10.1002/rmv.356

[43] Laskowski, R. A., Jabłońska, J., Pravda, L., Vařeková, R. S. and Thornton, J. M. PDBsum: Structural summaries of PDB entries. *Protein science*, 27: 129-134, 2018. https://doi.org/10.1002/pro.3289

[44] Hutchinson, E. G. and Thornton, J. M. HERA—a program to draw schematic diagrams of protein secondary structures. *Proteins: Structure, Function, Bioinformatics*, 8: 203-212, 1990. https://doi.org/10.1002/prot.340080303

[45] Krieger, E. and Vriend, G. YASARA View—molecular graphics for all devices—from smartphones to workstations. *Bioinformatics*, 30: 2981-2982, 2014. https://doi.org/10.1093/bioinformatics/btu426

[46] Land, H. and Humble, M. S. YASARA: a tool to obtain structural guidance in biocatalytic investigations. *Protein engineering: methods protocols*, 43-67, 2018. https://doi.org/10.1007/978-1-4939-7366-8_4

[47] Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22: 2577-2637, 1983. https://doi.org/10.1002/bip.360221211

[48] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. The Protein Data Bank. *Nucleic acids research,* 28: 235-242, 2000. https://doi.org/10.1093/nar/28.1.235

[49] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. and Ferrin, T. E. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry,* 25: 1605-1612, 2004. https://doi.org/10.1002/jcc.20084

[50] Ihaka, R. and Gentleman, R. R: a language for data analysis and graphics. *Journal of computational graphical statistics,* 5: 299-314, 1996. https://doi.org/10.1080/10618600.1996.10474713

[51] Love, M., Anders, S. and Huber, W. Differential analysis of count data–the DESeq2 package. *Genome Biology,* 15: 10-1186, 2014. http://dx.doi.org/10.1186/s13059-014-0550-8

[52] Li, B. and Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics,* 12: 1-16, 2011. https://doi.org/10.1186/1471-2105-12-323

[53] Cox, D. R. and Reid, N. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B,* 49: 1-18, 1987. https://doi.org/10.1111/j.2517-6161.1987.tb01422.x

[54] McCarthy, D. J., Chen, Y. and Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research,* 40: 4288-4297, 2012. https://doi.org/10.1093/nar/gks042

[55] Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W. and Zhan, L. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The innovation,* 2: 1-10, 2021. https://doi.org/10.1016/j.xinn.2021.100141

[56] Yu, G., Wang, L.-G., Han, Y. and He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology,* 16: 284-287, 2012. https://doi.org/10.1089/omi.2011.0118

[57] Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M. and Sherlock, G. GO:: TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics,* 20: 3710-3715, 2004. https://doi.org/10.1093/bioinformatics/bth456

[58] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S. and Eppig, J. T. Gene ontology: tool for the unification of biology. *Nature genetics*, 25: 25-29, 2000. https://doi.org/10.1038/75556

[59] Kanehisa, M. and Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28: 27-30, 2000. https://doi.org/10.1093/nar/28.1.27

[60] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R. and Lander, E. S. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102: 15545-15550, 2005. https://doi.org/10.1073/pnas.0506580102

[61] Huang, D. W., Sherman, B. T. and Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4: 44-57, 2009. https://doi.org/10.1038/nprot.2008.211

[62] Sherman, B. T., Huang, D. W., Tan, Q., Guo, Y., Bour, S., Liu, D., Stephens, R., Baseler, M. W., Lane, H. C. and Lempicki, R. A. DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC bioinformatics*, 8: 1-11, 2007. https://doi.org/10.1186/1471-2105-8-426