Chapter III

Materials and Methods

In this chapter we intended to describe the descriptors and methods which we employed in our work.

3.1: Calculation of molecular descriptors

Graph theoretical indices

A graph is the application of set itself i.e. a collection of elements of the set, and of binary relations between these elements. In case of a chemist the geometrical realization of graph is more appealing, namely a collection of points i.e. elements of the set. The name 'Graph' originates from this geometrical realization.

As the shape or length of the lines, or angles between lines has no specification so the graphs are topological rather than geometrical objects. Its most important features the vicinity relationship between points.

There are two types of correspondence between graphs and chemical categories in chemistry.

i) Structural or constitutional graphs—In this type molecules are presented by and covalent bonds are presented by lines.

ii) Reaction graphs—In this type chemical species are presented by points and the conversion between these species are presented by lines.

In our study we mainly deal with the structural or constitutional graphs only. The Graph theory was independently discovered on several occasions [1] and three names deserve special mention – Euler, Kirchoff and Cayley [2-4].

Euler published the first known paper on graph theory. It deals with Konigsberg bridge problem.

Kirchoff discovered graph theory while solving the calculation of currents in electrical networks [3]. Organic chemistry becomes the third breeding ground for graph theory. The early organic chemists Couper, Butlerov and Kekule founded the structure theory. They present a covalent bond between two atoms as line joining two points. In this way every structural formula is a graph. Cayleyput forward his concept of tree to enumeration of chemical isomers [4]. It was a challenging mathematical problem which was solved by him in 1874.

The graph theory has various applications in modern science like cryptography, networking etc.

The chemical structure of any chemical compound may be represented by graph and is termed as chemical graph. Characteristic invariants of graphs are related with structural property of molecules. These invariants are termed as topological indices.

Topological Indices:- The topological indices (TIs) are numerical invariants that quantitatively characterize molecular structure. A graph G=(V,E) is an ordered pair of two sets V and E. V represents a nonempty set and E represents unordered pair of elements of set V. When V represents the atoms of a molecule and elements of V symbolize covalent bonds between pairs of atoms, then G becomes a molecular graph. This type of graph also termed as constitutional graph, because there is no stereo-chemical information. A numerical graph invariant that characterizes the molecular structure is called a topological index. In this study we have calculated various topological indices like Information content, Structural Information content, etc. Some of the indices based on the nature of atom of its adjacent and some are depends on the bonds. Most of this can derive from the various matrices corresponding to a molecular graph.

Information theoretic topological indices are calculated by the application of information theory on chemical graphs. An appropriate set of n-elements is derived from a molecular graph G depending upon certain structural characteristics. On the basis of a equivalence relation defined on A1, the set A is partitioned into h disjoint subsets Ai of order (I=1, 2,... $h_i \sum_{i=1}^{n} n_i = n$). A probability distribution is then assigned to the set of equivalence classes.

A1, A2,A3,.....,Ah

p1, p2 .....ph

where $p_i = \frac{n_i}{n}$ is the probability that a randomly selected element of A will occur in the Ith subset.

The mean information content of an element of A is defined by Shanon's relation [11].

$$IC = -\sum_{i=1}^{n} p_i \log_2 p_i$$

The logarithm is taken at base 2 for measuring the information content in bits. The total information content of the set A is then n*IC.

In this method chemical species are symbolized by weighted linear graphs. Two vertices u0 and v0 of a molecular graph are said to be equivalent with respect to the rth order neighborhood if and only if ,corresponding to each path $u_1,u_2,....u_r$ of length r , there is a distinct path $v_1,v_2,.....v_r$ of the same length ,such that the paths have similar edge weights ,and both u_0 and v_0 are connected to the same number and type of atoms up to the rth order bonded neighbors.

Once partitioning of the vertex set for a particular order of neighborhood is completed, ICr is calculated from equation. Basak , Roy and Ghosh defined another information theoretic measure , Structural information content (SICr) [12] which is calculated as

$$SIC_r = \frac{IC_r}{\log_2 n}$$

Where ICr is calculated from equation and n is the total number of vertices of the graph.

Another information-theoretic invariant, Complementary information content (CICr) [13], is defined as

$$CIC_r = \log_2 n - IC$$

CICr represents the difference between the maximum possible complexity of a graph (where each vertex belongs to a separate equivalence class) and the realized topological information of chemical structures [14].

Dipole moment

The polarity of a molecule is represented by the dipole moment. Dipole moment can be defined as the product of magnitude of charge and the distance of separation between the charges. An electric dipole consists of a pair of charges of equal magnitude and opposite signs separate by a distance if the positive and negative charges in a molecule do not overlap, the molecule possesses a permanent dipole moment (μ) (polar molecule). Generally molecular dipole moment is calculated using the following formula.

$$\mu = \sum q_i r_i$$

Where r_i is the radius-vector of an atom i from the origin of the coordinate system (centre of charge or centre of mass) and q_i is the partial charge of atom i

The magnitude of one or more of the vector's components along the x, y and z Cartesian axes can also be used. Dipole moment is the measure of net molecular polarity of the molecule. The attraction between two polar molecules is called dipole-dipole interaction. The dipole moment produced electrostatic interactions with biological macromolecules. The energy of dipole-dipole interactions can be described by the following equation.

$$E = \frac{2\,\mu_1\mu_2 cos\theta_1 cos\theta_2}{Dr^3}$$

Where m is the dipole moment, u is the angle between the two poles of the dipole, D is the dielectric constant of the medium and r is the distance between the charges involved in the dipole.

The magnitude of the dipole moment is used as a descriptor in the QSAR analysis and the use of dipole moments in QSAR studies was proposed by Lien et al [15-17].

Polarizability

The polarizability of a molecule (α) is one of the most significant electrical properties, which characterizes the ability of the electron cloud of an atom or a molecule to be distorted from its normal shape by the external electric field. Due to this distortion of electronic system an induced electric dipole moment appears. It is defined by the coefficient of proportionality between the strength of an applied electric field (E) and the magnitude of the induced dipole moment (μ_{ind}) using the following equation.

$$\mu_{ind} = \alpha E$$

If a molecule have a small number of electrons, its polarizability is lower than that of the molecule containing atoms with a larger number of electrons and a more diffuse electron

distribution. Experimentally, polarizability is calculated by the Lorentz-Lorenz relation [18].

$$MR_D = \left(\frac{n_D^2 - 1}{n_D^2 + 2}\right)\frac{M}{\rho} = \frac{4}{3}\pi N_0 \alpha$$

Where, N_0 is the Avogadro constant, n_D is the refractive index, ρ is the dencity and M is the molecular mass.

The van der Waals (or London) forces are the universal attractive force between atoms that hold nonpolar molecules together in the liquid phase. This attractive force are base on polarizability and the fluctuating dipoles or shifts in electron clouds of the atoms tend to induce opposite dipoles in adjacent molecules, resulting in a net overall attraction. The energy of this interaction inversely proportion to $1/r^6$, where r is the distance separating the two molecules. The van der Waals force operates at a distance between 0.4–0.6 nm and exerts an attraction force less than 0.5 kcal/mol. although individual van der Waals forces make a low energy contribution to an event but they become significant when summed up over a large area with close surface contact of the atoms. The polarizability at the surface of both the drug and its binding site (receptor) contribute to the interaction energy. Abraham and coworkers introduced polarizability as parameters in QSAR [19, 20].

Refractivity

Refractivity (MR) one of the most important chemico-physical properties used as descriptor in QSAR studies. It has been shown to be related to lipophilicity, molar volume and steric bulk. The importance of splitting the MR into its atomic components for QSAR studies oriented to three-dimensional molecules was demonstrated by Crippen et al [21].

Partition coefficient (logP)

Partition coefficient (P) is defined as the ratio of concentration of the solute in the organic phase to the non-ionised solute concentration in the water phase, at equilibrium.

$$P = \frac{C_{org}}{C_{water}}$$

Where C_{org} is the equilibrium concentration of the solute in the oil phase and C_{water} is the equilibrium concentration of the solute in water [22]. P denote the distribution of a compound between two phases organic and water. Partition coefficient used in its logarithmic form (logP). The zero logP value indicates that the solute is equally soluble in the two phases, a negative logP imply that the solute is more soluble in water, and a positive value suggests that a greater solubility in the oil phase. Examples of organic phases are octanol, cyclohexane and chloroform etc. LogP is probably the most commonly used descriptor of lipophilicity or hydrophobicity and it is generally understand the ability of the solute to cross lipid membranes. Lipophilicity involves many stages of drug action. Prior to reaching a pharmacological target, lipophilicity determines solubility, reactivity and degradation of drugs, as well as formulation of pharmaceuticals. Moreover, compound lipophilicity is of principal importance for biological activity as the affinity for a lipophilic environment facilitates the transport of chemicals through membranes in a biological system and the formation of complexes between compounds and receptor binding site. Cell membrane, a selectively- permeable barrier, mostly consists of a phospholipid bilayer with embedded proteins. Amphiphilic phospholipids composed of fatty acid chains at one end and hydrophilic ionized head regions at the other arrange spontaneously in the lipid bilayer. The drug interaction with lipid structures present in the organism is strongly related to its lipophilicity. Molecules with lower logP values cannot easily enter the lipid phase of the membranes, whereas molecules

with high logP values are trapped in the membrane. Therefore molecules have intermediate logP values (e.g. between about 0 and4) can readily cross cell membranes [23, 24].

LogP can be calculated by a number of experimental methods. However, various computational methods and software are available for calculation of logP for the octanol-water system.

van der Waals surfaces area (VSA) and solvent-accessible surface area (SASA)

Molecular surface area is the area of the outer surface of the volume from which solvent molecules are excluded due to the presence of the solute molecule in a solution. It is based on the Van der waals molecular surface (defined by the Van der Waals radii of the atoms (represented as spheres) in the molecule), however, Van der waals molecular surface contains small gaps and crevices, which are inaccessible to other atoms and molecules (for example solvent molecules). The molecular surface area is defined by excluding these gaps and crevices. Thus, the molecular surface consists of the Van der waals surfaces of the atoms where they can enter in a contact with the solvent molecules, and additionally, of the surfaces of the solvent molecules, placed in contacts with the Van der waals surfaces of two or more atoms of the investigated molecule [25].

Typically water is used as a solvent to perform calculations of the molecular surface area. For practical reasons the shape of the water molecule is considered as a sphere with a radius of 1.4 to 1.7 Å, which is the average distance from the centre of the oxygen atom to the Van der aals surface of the water molecule. A solvent-accessible molecular surface area is defined by the centre of a probe sphere (solvent molecule, typically water), when it is rolled over the molecular surface.

Energies of the frontier molecular orbital's HOMO and LUMO

Partial atomic charges appear due to the different ability of atoms to withdraw electron The energies of the frontier orbitals HOMO (highest occupied molecular orbital) and LUMO (lowest unoccupied molecular orbital) are very popular quantum chemical descriptors in QSAR. These orbitals play an important role in many chemical reactions and determining electronic band gaps in solids [26]. These orbitals are also responsible for the formation of many charge transfer complexes [27, 28]. According to the frontier molecular orbital theory of chemical reaction, transition state formation involves the interaction between the frontier orbitals (HOMO and LUMO) of reacting species. Thus, the treatment of the frontier molecular orbitals separately from the other orbitals is based on the nature of chemical reactions [29]. The HOMO energy is directly related to the ionization potential and characterizes the susceptibility of the molecule toward attack by electrophiles, where as the energy of the LUMO is directly related to the electron affinity and characterizes the susceptibility of the molecule toward attack by nucleophiles. A higher HOMO energy imply higher affinity of a molecule to react as a nucleophile and a lower LUMO energy suggests stronger electrophilic nature of a molecule. The difference in energy between the HOMO and the LUMO is an important for the stability. A large HOMO-LUMO gap implies high stability for the molecule i.e. the lower reactivity in chemical reactions The HOMO and LUMO energies are calculated with the methods of the quantum mechanics [30,31].

Commonly used software for calculating Molecular Descriptors are HyperChem (HyperChem 5.1, Hypercube Inc., Gainesville, Florida, USA), ACD/LogP KOWWIN, Pallas, TOPKAT, Dragon (TALETE srl), GAMESS, MOPAC, MervinlogP calculator,

3.2 Docking simulation methods

With the rapid increase in computational power, in silico methods became widely used in the fields of structural molecular biology and structure-based drug design. Molecular docking is one of these computational techniques [32, 33]. Docking is a method involves the prediction of the preferred orientation of one molecule to second which bound to each other to form a stable complex. Preferred orientation of the molecule may be used to predict the strength of association or binding affinity between two molecules using for example scoring functions. Docking is generally devised as a multi-step process in which each step introduces one or more additional degrees of complexity. The process begins with the application of docking algorithms that fit the small molecules in the active site of substrate.

Examples of some popular protein-ligand docking systems include AutoDock [34], GOLD [35], DOCK [36], GLIDE [37], ICM [38], FlexX [39] and SITUS [40]. Of these, molecular docking simulation was carried out using the autodock 4.2. AutoDock has verified software capable of quickly and accurately predicting bound conformations and binding energies of ligands with protein.

AutoDock employed a grid-based method, which permit rapid evaluation of the binding energy of trial conformations .Grid maps are calculated by AutoGrid. AutoDock need pre-calculated grid maps. Grid maps of each atom type present in the ligand being docked. Grids maps help to make the docking calculations extremely fast. A grid map consists of a three dimensional lattice of regularly spaced points, surrounding and centered on some region of interest of the macromolecule under study. Typical grid point spacing varies from 0.2 Å to 1.0 Å (default value is 0.375 Å). Each point within the grid map stores the potential energy of a probe atom or fictional group that is due to all the atoms in the macromolecule.

Morris et al. consider three search methods, the Lamarckian genetic algorithm, Monte Carlo simulated annealing, anda traditional genetic algorithm. The primary method for conformational searching is a Lamarckian genetic algorithm [41]. A population of trial conformations is created, and then in successive generations these individuals mutate, exchange conformational parameters, and compete in a manner analogous to biological evolution, ultimately selecting individuals with lowest binding energy. The "Lamarckian" aspect is an added feature that allows individual conformations to search their local conformational space, finding local minima, and then pass this information to later generations. The Lamarckian Genetic Algorithm provides the most efficient search for general applications and in most cases will be the technique used. It is typically effective for systems with about 10 rotatable bonds in the ligand. A Lamarckian genetic algorithm combined with a scoring function based on the AMBER force field [42].

MonteCarlo (MC) methods are amongthe most established and widely used stochastic optimization techniques [43]. The combination of atomistic potential energy models with stochastic search techniques has produced some of the most powerful methods for both structure optimization and prediction.

The docking simulation is carried out using the Metropolis method, also known as Monte Carlo simulated annealing. With the protein static throughout the simulation, the substrate molecule performs a random walk in the space around the protein. At each step in the simulation, a small random displacement is applied to each of the degrees of freedom of the substrate: translation of its center of gravity; orientation; and rotation around each of its flexible internal dihedral angles.

This displacement results in a new configuration, whose energy is evaluated using the grid interpolation procedure described above. This new energy is compared to the energy

of the preceding step. If the new energy is lower, the new configuration is immediately accepted. If the new energy is higher, then the configuration is accepted or rejected based upon a probability expression dependent on a user defined temperature, T. The probability of acceptance is given by

$$P(\Delta) = e^{(-\frac{\Delta E}{k_B T})}$$

Where ΔE is the difference in energy from the previous step, and kB is the Boltzmann constant. At high enough temperatures, almost all steps are accepted. At lower temperatures, fewer high energy structures are accepted.

The simulation proceeds as a series of cycles, each at a specified temperature. Each cycle contains a large number of individual steps, accepting or rejecting the steps based upon the current temperature. After a specified number of acceptances or rejections, the next cycle begins with a temperature lowered by a specified schedule such as

$$T_i = gT_{i-1}$$

Where Ti is the temperature at cycle i, and g is a constant between 0 and 1.

Simulated annealing allows an efficient exploration of the complex configurational space with multiple minima that is typical of a docking problem. The separation of the calculation of the molecular affinity grids from the docking simulation provides modularity to the procedure, allowing the exploration of a range of representations of molecular interactions, from constant dielectrics to finite difference methods and from standard 12-6 potential functions to distributions based on observed binding sites.

The Genetic Algorithm may also be run without the local search, but this is typically less efficient than the Lamarckian genetic algorithm combination.

AutoDock4 uses a semiempirical free energy force field to predict binding free energies of small molecules to macromolecular targets. The semiempirical free energy force field estimates the energetic of the process of binding of two (or more) molecules in a water environment using pair-wise terms to evaluate the interaction between the two molecules and an empirical method to estimate the contribution of the surrounding water. This differs from a traditional molecular mechanics force field, which also relies on pair-wise atomic terms, but typically uses explicit water molecules to evaluate solvation contributions. The goal of the empirical free energy force field is to capture the complex enthalpic and entropic contributions in a limited number of easily evaluated terms.

The protein and ligand start in an unbound conformation. The force field evaluates the free energy of binding in two steps. The free energy of binding is estimated to be equal to the difference between (1) the energy of the ligand and the protein in a separated unbound state and (2) the energy of the ligand and protein in their bound conformation (complex). Evaluate the intramolecular energetics of the transition from the unbound state to the bound conformation for each of the molecules separately and then evaluate the intermolecular energetic of bringing the two molecules together into the bound complex.

The force field includes six pair-wise evaluations (V) and an estimate of the conformational entropy lost upon binding (ΔS_{conf}):

$$\Delta G = (V_{bound}^{L-L} - V_{unbound}^{L-L}) + (V_{bound}^{P-P} - V_{unbound}^{L-L}) + (V_{bound}^{P-L} - V_{unbound}^{P-L} + \Delta S_{conf})$$

In equation 1, L refers to the "ligand" and P refers to the "protein" in a protein–ligand complex; this approach is equally applicable for any types of molecules in a complex. The first two terms are intramolecular energies for the bound and unbound states of the ligand, and the following two terms are intramolecular energies for the bound and unbound states of the protein. The third parentheses represent the change in

intermolecular energy between the bound and unbound states of protein and ligand. It is assumed that the two molecules are sufficiently distant from one another in the unbound state that $V_{unbound}^{P-L}$ is zero. The bound state of the protein is identical with the protein unbound state, and the difference in their intramolecular energy is zero.

The pair-wise atomic terms include evaluations for dispersion/repulsion, hydrogen bonding, electrostatics, and desolvation:

$$V = W_{vdw} \sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + W_{hbond} \sum_{i,j} E(t)\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right) + W_{elec} \sum_{i,j} \frac{q_i q_j}{\epsilon r_{ij} r_{ij}} +$$

$$W_{sol} \sum_{i,j}(S_i V_j + S_j V_i) e^{-\frac{r_{ij}^2}{2\sigma^2}}$$

The weighting constants W are the ones that are optimized to calibrate the empirical free energy based on a set of experimentally determining binding constant. The first term is a typical 6/12 potential for dispersion/repulsion interactions. The Parameters A and B were taken from the Amber force field [44]. The second term is a directional H-bond term based on a 10/12 potential [45]. The parameters C and D are assigned to give a maximal well depth of 5 kcal/mol at 1.9 A° for hydrogen bonds with oxygen and nitrogen and a depth of 1 kcal/ mol at 2.5 A° for hydrogen bonds with sulpher [46,47]. Directionality of the hydrogen bond interaction represented by the function E(t), is dependent on the angle t away from ideal bonding geometry [48]. Directionality is further enhanced by limiting the number of hydrogen bonds available to each point in the grid to the actual number of hydrogen bonds that could be formed. The third term is a screened Coulomb potential for electrostatic interactions [49]. The final term is a desolvation potential based on the volume (V) of the atoms surrounding a given atom, weighted by a solvation parameter (S) and an exponential term based on the distance. The distance weighting factor σ is set to 3.5 A° [50].

This force field is standardized for a united atom model, which explicitly includes heavy atoms and polar hydrogen atoms. Intramolecular energies are calculated for all pairs of atoms within the ligand (or protein, if it has free torsional degrees of freedom), excluding 1–2, 1–3, and 1–4 interactions.

The term for the loss of torsional entropy upon binding (ΔS_{conf}) is directly proportional to the number of rotatable bonds in the molecule (N_{tors}):

$$\Delta S_{conf} = W_{conf} N_{tors}$$

The number of rotatable bonds includes all torsional degrees of freedom, including rotation of polar hydrogen atoms on hydroxyl groups and the like.

Desolvation

In the development of an empirical free energy function for AUTODOCK, the desolvation term was most challenging, because AUTODOCK uses a grid based method for energy evaluation, and most published solvation methods are based on surface area calculations.

The desolvation term is calculated using the general approach of Wesson and Eisenberg [51]. For the Calculate of desolvation energy, two information are needed (1) an atomic solvation parameter for each atom type, which is an estimate of the energy needed to transfer the atom between a fully hydrated state and a fully buried state and (2) an estimate of the amount of desolvation when the ligand is docked. The amount of desolvation is calculated by the volume-summing method, which is similar to the Stouten et al. method [50]. Huey et al. introduced a modified approach for the atomic solvation parameters based on the chemical type and the atomic charge of the atom. This approach employed in AutoDock and other docking methods. Incorporation of the atomic charge

into the salvation parameter removes the need to use two discrete charged and uncharged atom types for oxygen and nitrogen.

The solvation parameter (S_i) for a given atom is calculated as:

$$S_i = (ASP_i + QASP \times |q_i|)$$

Where q_i refers to the atomic charge and ASP and QASP are the atomic solvation parameters derived here. The ASP is calibrated using six atom types such as aliphatic carbons (C), aromatic carbons (A), nitrogen, oxygen, sulfur, and hydrogen. A single QASP is calibrated over the set of charges on all atom types.

For each atom in the protein, the volume term in the free energy equation was evaluated:

$$\Delta V_i = \sum_{k \neq i} V_k \times e^{-\frac{k}{2\sigma^2}}$$

Where k is all atoms in the protein and all atoms in the same amino acid residue represented as i. The maximal value of ΔV for each amino acid type over the entire set of proteins was then determined. These values were used to perform a least-squares fit of the model to a set of experimental vacuum-to-water transfer energies to determine values for the atomic solvation parameters ASP and QASP [52].

We used a simple approximation for incorporation of additional atom types in the desolvation model. The ASP is assigned to the average of the values from the six atom types used in the calibration and the same QASP is applied.

Unbound States

In order to estimate a free energy of binding, AutoDock needs to estimate energy for the unbound state of the ligand and protein. Morris et al. investigated three approaches to the

unbound state. The first approach is a fully extended conformation, which models a fully solvated conformation with few internal contacts. A short optimization was performed on the ligand in isolation using a uniform potential inversely proportional to the distance between each pair of atoms. This moves all atoms as far away from one another as possible. The second approach is a minimized conformation that has substantial internal contacts, modeling a folded state for the unbound ligand. A short Lamarckian genetic algorithm conformational search was performed, using an empty affinity grid. As expected, these conformations tend to bury hydrophobic portions inside and form internal hydrogen bond interactions.

The third approach the assumption used in AutoDock3 and many other docking methods. In this, it is assumed that the conformation of the unbound state is identical to the conformation of the bound state.

Coordinate Sets

The force field was standardized and tested using a large collection of protein complexes for which experimental information on binding strength is available. The force field was calibrated on a set of 188 complexes. Binding data were collected from the Ligand–Protein Database (http://lpdb.scripps.edu), and coordinates were obtained from the Protein Data Bank (http://www.pdb.org). These complexes were checked and corrected if necessary. Hydrogen atoms were added automatically using Babel, atomic charges were added using the Gasteiger PEOE method, and then nonpolar hydrogen atoms were merged [53-54]. The Gasteiger method was selected for its fast and easy operation and ready availability as part of Babel. Ligands were processed in ADT (http://autodock.scripps.edu/resources/adt) to assign atom types and torsion degrees of freedom. Finally, a short optimization of the ligand was performed.

Redocking

Redocking experiments were carried out with AutoDock4 and the new empirical free energy force field. For each complex, 50 docking experiments were performed using the Lamarckian genetic algorithm conformational search with the default parameters from AutoDock3. A maximum of 25 million energy evaluations was applied for each experiment. The results were clustered using a tolerance of 2.0 A°.

3.3: Molecular dynamics simulation methods

Molecular dynamics is an important tool to investigate the microscopic behaviors by integrating the motions of particles or particle clusters, based on Newtonian dynamics. Theoretically it has been applied to study the dynamics of a macromolecular system. Now molecular dynamics simulations have been used most widely for studying protein motions.

At the commencement of a dynamic simulation, an initial set of atomic coordinates and velocities are needed. Generally, these are obtained from the X-ray crystallographic or NMR structure data, or by model-building (based on the structure of a homologous protein, for example). Given a set of coordinates, a preliminary calculation serves to equilibrate the system.

The structures are first refined using an energy minimization algorithm to reduce local stresses due to the nonbonded atomic overlaps, distortions of bond length, etc.

Then the protein atoms are assigned velocities (v) taken at random from a Maxwellian distribution corresponding to a low temperature, and a simulation is performed for a period of a few picoseconds.

This is done bythe laws of classical mechanics, and most notably the Newton's 2nd law:

$$F_i = m_i a_i$$

Where F_i the force acting upon ith particle at the time, m_i is the atomic mass of the particle, a_i is the acceleration ($a_i = d^2r_i/dt^2$) and introducing it into the equation for the position r_i at time $t+\Delta t$, given r_i at time t:

$$r_i(t+\Delta t) = r_i(t)+v_i\Delta t+\frac{1}{2}a_i(\Delta t)^2$$

The equilibration is continued by alternating new velocity assignments, chosen from Maxwellian distributions corresponding to successively increased temperatures with intervals of dynamical relaxation. The temperature T is measured in terms of the mean kinetic energy for the system composed of N atoms:

$$\frac{1}{2}\sum_{i=1}^{N} m_i v_i^2 = \frac{3}{2} N k_B T$$

Where v^2_i *is the* average velocity squared of the ith atom and k_B is the Boltzmann constant. The equilibration period is considered completed when the temperature is stable for longer than about 10 ps, the atomic momenta obey maxwellian distribution and different regions of the protein have the same average temperature.

Integration of the equations of motion after equilibration generates the coordinates and velocities of the atoms as a function of time. Many numerical algorithms are used to solve the equation of motion such as Predictor-corrector algorithm, Verlet algorithm and Leap-frog algorithm etc.

Predictor corrector algorithm

Predictor-corrector algorithms compose commonly used class of methods to integrate the equations of motion and more frequently used in molecular dynamics due to Gear, and consists of three steps, namely :Predictor, Force evaluation, Corrector [55].

If the classical trajectory is continuous then an estimate of the positions, velocities etc. at time t+∂t may be obtained by Taylor expansion about time t:

$$r^p(t+\partial t) = r(t) + v(t)\partial t + \frac{1}{2}a(t)\partial t^2 + \ldots\ldots$$

$$v^p(t+\partial t) = v(t) + a(t)\partial t + \frac{1}{2}b(t)\partial t^2 + \ldots\ldots$$

$$a^p(t+\partial t) = a(t) + b(t)\partial t + \ldots\ldots$$

Where r represents the position, v is the velocity (the first derivative with respect to time), a is the acceleration (the second derivative with respect to time), etc. But the above equation will not generate correct trajectories as time advances and not introduced the equation of motion. These enter through the correction step. Calculated the new position r^p, the forces at time t+∂t and hence the correct accelerations a^c (t+∂t). These can be compared with the predicted acceleration from the above equation.

To estimate the size of the error in the prediction step:

$$\Delta a(t+\partial t) = a^C(t+\partial t) - a^P(t+\partial t)$$

This error and the results of predictor step are fed into the corrector which gives;

$$r^C(t+\partial t) = r^P(t+\partial t) + c_o \Delta a(t+\partial t)$$

$$V^C(t+\partial t) = v^P(t+\partial t) + c_1 \Delta a(t+\partial t)$$

$$a^c(t+\partial t) = a^P(t+\partial t) + c_2 \Delta a(t+\partial t)$$

The idea is that $r^c(t+\partial t)$ etc are now better approximation to the true positions, velocities etc. The general scheme of a stepwise MD simulation based on a predictor-corrector algorithm, which may be summarized as follow:

(a) Predict the positions velocities accelerations at time $(t+\partial t)$ using the correct values of these equation.

(b) Evaluate the forces and hence accelerations $a_i = f_i/m_i$, from the new position.

(c) Correct the predicted positions velocities accelerations using the new acceleration.

(d) Calculate any variables of interest such as energy, order parameters before returning to a for the next step.

Verlet algorithm

One of the most simplest and common method of integrating the equation of motion is called Verlet algorithm [56-57]. The method is a direct solution of the second order equations. The Verlet algorithm uses positions and accelerations at time t and the positions from time $t-\partial t$ to calculate new positions at time $t+\partial t$. The Verlet algorithm uses no explicit velocities.

$$r(t + \partial t) = r(t) + v(t)\partial t + \tfrac{1}{2}a(t)\partial t^2$$

$$r(t - \partial t) = r(t) - v(t) + \tfrac{1}{2}a(t)\partial t^2$$

Summing these two equations, one obtains

$$r(t + \partial t) = 2r(t) - r(t - \partial t) + a(t)\partial t^2$$

The velocities do not explicitly appear in Verlet algorithm but they are useful for calculating the kinetic energy and hence the total energy. They may be obtained from the formula,

$$V(t) = \{ r(t+\partial t) - r(t-\partial t)\}/2\partial t$$

The velocities are not required to compute the trajectories, but they are useful for calculating observables like the kinetic energy. Success of the Verlet algorithm is straightforward and also storage requirements are modest, comprising two sets of positions (r(t) and r(t-δt)) and accelerations(a(t)).

Leap-frog algorithm

Leap-frog integration is equivalent to calculating positions and velocities at interleaved time points, interleaved in such a way that they "leapfrog" over each other. In this algorithm, the velocities are first calculated at time *t+ 1/2dt;* these are used to calculate the positions, *r,* at time *t+dt.* In this way, the velocities leap over the positions, and then the positions *leap* over the velocities [58-59].

$$r(t + \partial t) = r(t) + v(t + \tfrac{1}{2}\partial t)\partial t$$

$$v(t + \tfrac{1}{2}\partial t) = v(t - \tfrac{1}{2}\partial t) + a(t)\partial t$$

The advantage of this algorithm is that the velocities are explicitly calculated, however, the disadvantage is that they are not calculated at the same time as the positions. The velocities at time t can be approximated by the relationship:

$$v(t) = 1/2[v\{t-½\partial t) + v(t+½\partial t)]$$

In our study MD simulation is performed using GROMACS software.

Different types of software are used in molecular dynamics simulation. Some common and widely used software are given in appendix III.

3.4: References

1. Hrary F (1969) Graph theory. Chapter 1, Addision Wesley, Reading mass
2. Euler. L. (1736) Comment. Acad. Sci. I. Petropolitanae, 8, 128, translated in Sci. Amer, 189, 66 (1953)
3. Kirchoff. G. (1842) Ann. Phys. Chem. 72,497.
4. Cayley A. (1874). Philos. Mag. 67, 444.
5. M. Randic, P.J. Hansen and P.C. Jurs, J. Chem. Inf. Comput. Sci. 28 (1988) 60.
6. R. Todeschini and V. Consonni, Handbook of Molecular Descriptors (Wiley-VCH, Weinheim, 2000).
7. M. Randic, On characterization of molecular branching, J. Amer. Chem. Soc. 97 (1975) 6609.
8. Wiener HJ (1947) *Am. Chem. Soc.*, **69**, 17-20
9. Harary F (1969) Graph theory. Addison-Wesley, Reading MA.

10. Balaban AT, Harary F (1968) Chemical graphs. V. Enumeration and proposed nomenclature of benzenoidcata condensed polycyclic aromatic hydrocarbons. *Tetrahedron*. **24**, 2505–2516.

11. Shanon CE (1948) A mathematical theory of Communication. *Bell Syst. Tech.J.*, *27*: 379-423.

12. Basak SC, Roy AB, Ghosh JJ (1979) Study of Structure-Function Relationship of Pharmacological and Toxicological Agents Using Information Theory. In proceeding of the llnd International conference on mathematical modeling, University of Missouri at *Rolla*. 851-856.

13. Basak SC and Magnuson VR (1983) Molecular Topology and Narcosis. A Quantitative Structure-Activity Relationship (QSAR) Study of Alcohols Usins Complementary Information Content, Arzneimittle-Froschung, *Drug Research*. 501-503.

14. Lien EJ, Liao RCH, Shinouda HGJ (1979) *Pharm. Sci.*, 68: 463-465.

15. Lien EJ, Guo ZR, Li RL, Su CTJ (1982)*Pharm. Sci.* **71**: 641-655.

16. Li WY, Guo ZR, Lien EJJ (1984) *Pharm. Sci.* **73**: 553-558.

17. Kurt E. Oughstun and Natalie A (2003) Cartwright. On the Lorentz-Lorenz formula and the Lorentz model of dielectric dispersion. OPTICS EXPRESS. **11:** 1541.

18. Abraham MH (2004)The factors that influence permeation across the blood-brain barrier. *European Journal of Medicinal Chemistry*, **39**:235–240.

19. Gaudio AC, Korolkovas A,Takahata Y (1994) Quantitative structure activity relationships for 1,4-Dihydropyridine calcium channel antagonists (nifedipine analogs) - a quantum chemical classical approach. *Journal of pharmaceutical sciences*, **83**: 1110–1115.

20. Ghose AK, Crippen GM (1987) Atomic Physicochemical Parameters for Three-Dimensional-Structure-Directed Quantitative Structure-Activity Relationships. 2. Modeling Dispersive and Hydrophobic Interactions. *J. Chem.ln$Comput.Sci.*,**27**: 21-35.

21. Smith RN, Hansch C, Ames MM (1975)*J. Pharm. Sci.*, **64**: 599.

22. Tomlinson E, David SS, Parr GD, James M, Farraj N, Kinkel JFM, Gaisser D, Wynn HJ in W. J. Dunn III, J. H. Block, and R. S. Pearlman, Eds., Partition Coefficient, Determination and Estimation, Pergamon, Oxford, UK, 1986, p. 83.

23. Hansch C, Kerley R (1970) Role of the Benzyl Moiety in Biochemical and Pharmacological Processes. *J. Med. Chem.*,**13**: 957-964.

24. Iwamura H (1980) Structure-Taste Relationehip of Perillartrine and Nitro and Cyanoaniline Derivatives. *J. Med. Chem.*, **23**: 308-312. D. E. Clark. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood-brain barrier penetration. Journal of Pharmaceutical Sciences, 88(8):815–821, 1999.

25. Zhou Z, Parr RGJ (1990) Am. Chem. Soc., 112, 5720.

26. Franke R (1984) Theoretical Drug Design Methods; Elsevier: Amsterdam. 115-123.

27. Osmialowski K, Halkiewicz J, Radecki A, Kaliszan RJ (1985).*Chromatogr.*, 346, 53.

28. Fukui K (1975) Theory of Orientation and Stereoselection; Springer Verlag: New York. 34-39.

29. Sklenar H, Ja¨ger J (1979)*Int. J. Quantum Chem.*, **16**: 467.

30. Fleming I (1976) Frontier Orbitals and Organic Chemical Reactions; John Wiley & Sons: New York.

31. Lengauer T,Rarey M, (1996) Computational methods for biomolecular docking. *Curr. Opin. Struct. Biol.*, **6**:402–406.

32. Gschwend DA, Good AC, Kuntz ID (1996) Molecular docking towards drug discovery. *Journal of Molecular Recognition*, **9**:175–186.

33. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK,Olson AJ(1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, **19**: 1639–1662.

34. Jones G, Willett P, Glen RC (1995) Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal of Molecular Biology*, **245**:43–53.

35. Shoichet BK, Kuntz ID, Bodian DL (1992) Molecular docking using shape descriptors. *Journal of Computational Chemistry*, 13:380–397.

36. Repasky MP, Shelley M, Friesner RA (2007) Flexible ligand docking with Glide. *Current Protocols in Bioinformatics*, 8.

37. Abagyan R, Totrov M,Kuznetsov D (1994) ICM - a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *Journal of Computational Chemistry*, **15**: 488–506.

38. Rarey M, Kramer B, Lengauer T,Klebe G () A fast flexible docking method using an incremental construction algorithm, *J Mol Biol.*,**261**: 470--489,.

39. Chacon P, Wriggers W (2002) Multi-Resolution Contour-Based Fitting of Macromolecular Structures. *Journal of Molecular Biology*, **317**: 375–384.

40. Morris GM, Goodsell DS, Halliday RS, HueyR, Hart WE, Belew RK, Olson AJJ (1998) *Comput Chem.*, **19**: 1639.

41. Cornell WD, Cieplak P,Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW,Kollman PA (1995)A SecondGeneration Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, **117**, 5179–5197.

42. Hart TN, Read RJ (1994) Multiple-start Monte Carlo docking of flexible ligands. Birkhaeuser, Boston.

43. Weiner SJ, Kollman PA,Case DA, Singh UC, Ghio C, Alagona G, Profeta S, Weiner PJ (1984)*Am Chem Soc.*, **106**: 765.

44. Morris GM, Goodsel DS, Halliday R, Huey R, Hart WE, Belew RK, Olson AJJ (1998) *Comput Chem.*, **19**, 1639.

45. Boobbyer DNA,Goodford PJ, McWhinnie PM, Wade RCJ (1989)*Med Chem.*, *32*: 1083.

46. Huey R,Goodsell DS, Morris GM, Olson AJ (2004)*Lett Drug Des Discov.*, **1**: 178.

47. Mehler EL, Solmajer T (1991) *Protein Eng.*,**4**, 903.

48. Stouten PFW, Frommel C, Nakamura H, Sander C (1993)MolSimul.,**10**, 97.

49. Wesson L, Eisenberg D (1992)*Protein Sci.*, **1**, 227.

50. Sharp KA, Nicholls A, Friedman R, Honig (1991)*Biochemistry*, **30**: 9686.

51. Tetko IV, Gasteiger J, Todeschini R., Mauri A, LivingstoneDJ, et al. (2005)*Comput Aided Mol Des*, **19**: 453.

52. Gasteiger J, Marsili M (1980)*Tetrahedron*, **36**: 3219.

53. C Gear (1971) Numerical Initial Value Problems in Ordinary Differential Equation, Prentice-Hall, Englewood Cliffs, NJ.

54. Allen M, Tildesley (1987) Computer Simulation of Liquids, Clarendon Press, New, York.

55. Frenkel D, SmitB, (2002) "Understanding molecular simulation: from algorithms to applications," 2nd ed., Academic Press, San Diego.

56. Gunsteren v, WF, Berendsen, H. J. C. A leap-frog algorithm for stochastic dynamics. Mol. Sim. 1:173–185, 1988.

57. Hockney RW, Goel SP, Eastwood J (1974) Quiet High Resolution Computer Models of Plasma. J. Comp. Phys. 14:148–158.