CHAPTER 2
THEORY

2.1 Molecular modeling

Molecular modeling was considered as a way to mimic the behavior of the molecules and molecular systems by using a simplified system which is often in term of mathematical terms to assist the calculations and predictions. The modeling today is associated with the computer modeling.

2.1.1 Homology modeling

Protein and nucleic acid sequences are determined routinely in molecular biology laboratories. The sequences are then generated rapidly and stored in many central databases such as EMBL database, PIR/NBRF database, GenBank database. On the other hand, the three dimensional structure of either protein or DNA determined by x-ray crystallography and NMR studies are obtained slower than that of their sequences. This made structural data of many proteins are not known. The methods in which the character of a protein can be predicted from its amino acid sequences were developed, i.e., building a protein by homology used only one known structure **(65)**, determining the protein structure using more than one references proteins **(66)**.

The homology model building process can be divided into the following steps **(67)**:

1. Determination of which proteins are related to the model protein.
2. Determination of structurally conserved region (SCRs).
3. Alignment of the amino acid sequences between the unknown protein and the reference protein(s) within the SCRs.
4. Assignment of the coordinates in the conserved regions.
5. Prediction of the conformation for the rest of the peptide chain, including loops between the SCRs and the N- and C-termini.
6. Investigation of the optimum side chain conformations for residues that differ from

those in the reference proteins.

7. Refinement of the molecular structure using energy minimization and molecular dynamics to relieve the steric strain introduced during the model-building process.

2.1.1.1 Sequence alignment

The Needleman and Wunsch algorithm **(68)** was used to align two sequences to identify the correspondence regions. The scores were given for the pairwise alignment between corresponding amino acids in the sequences, high scores for good matches. As for regions appear when the sequences are not of the same length, the gaps are introduced and given negative scores. The optimum alignment is led by the balance between the number of good matches and the least number of required gaps. The procedure for alignment is as follow:

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASN	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ASP	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CYS	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLN	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GLU	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
GLY	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
HIS	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ILE	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
LEU	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
LYS	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
MET	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
PHE	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
PRO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
SER	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
THR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
TRP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
TYR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
VAL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure 2.1 Amino acid identity matrix **(67)**.

Step 1 Set up a comparison matrix between the two sequences.

The matrix, which dimension equals to the lengths of the two sequences, has an element taken directly from one of scoring matrices. The identity matrix, one of the scoring

matrices gives a score of 1 for identical matches and 0 for all nonidentical pairs, is shown in Figure 2.1.

Step 2 Find the maximum pathway through the comparison matrix.

Each SCR regions were treated independently, no gaps are allowed within the conserved regions. In this case, an unknown sequence is added to the reference sequence as shown in Figure 2.2. A partial comparison is made between the unknown sequence and the reference sequence by using only the amino acids within the SCR regions (see Figure 2.3). The comparison gives the value taken from the identity matrix.

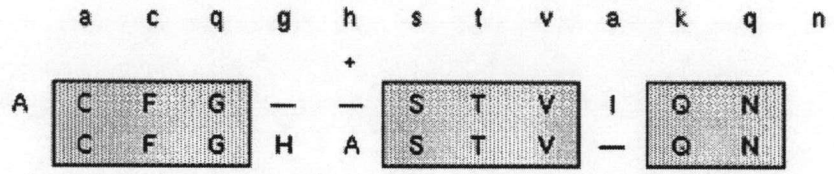

Figure 2.2 Addition of an unknown sequence to the alignment, the unknown sequence is shown in lowercase letters; boxes are drawn around the conserved regions of the aligned pair of sequences **(67)**.

	C	F	G	S	T	V	Q	N
a	0	0	0	0	0	0	0	0
c	1	0	0	0	0	0	0	0
q	0	0	0	0	0	0	1	0
g	0	0	1	0	0	0	0	0
h	0	0	0	0	0	0	0	0
s	0	0	0	1	0	0	0	0
t	0	0	0	0	1	0	0	0
v	0	0	0	0	0	1	0	0
a	0	0	0	0	0	0	0	0
k	0	0	0	0	0	0	0	0
q	0	0	0	0	0	0	1	0
n	0	0	0	0	0	0	0	1

Figure 2.3 A partial comparison matrix between the unknown sequence (lowercase letters listed along the left side) and the reference sequence (uppercase letter listed across the top) **(67)**.

After mapping between the unknown and the reference sequences, gaps can be placed in the proper locations. The pathway through the matrix is mapped and the gap regions were inserted between the SCR regions. The number of gaps inserted between the SCR regions is equal to the number of rows skipped (see Figure 2.4).

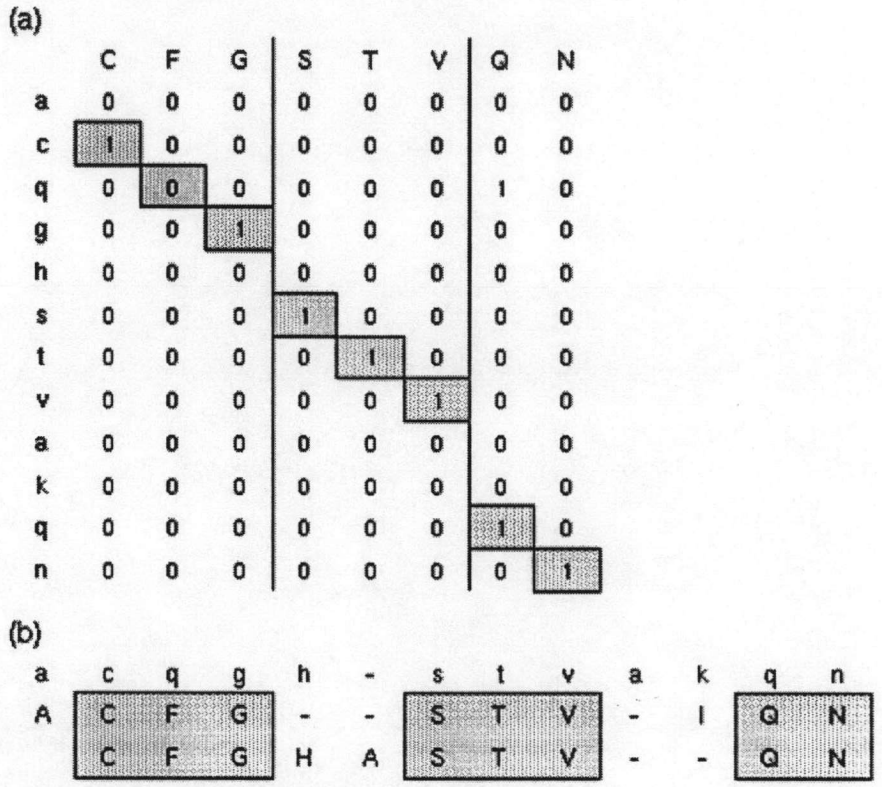

Figure 2.4 Mapping the pathway though the matrix (a) the final path through the comparison matrix, and (b) the final sequence alignment **(67)**.

2.1.1.2 Assignment of coordinates

The coordinates of the SCR regions can be assigned when the correspondence between amino acids in the reference and model sequences has been made **(67)**. The coordinates of the reference proteins are copied and transformed into the same coordinate frame for the model. Using the fact that in case of the side chains of the model and the reference are in the same locations along the sequence of the SCR, all coordinates for the amino acids are transferred. On the other hand, in case they are differ,

only the backbone coordinates are transferred. The side chain atoms are automatically replaced to preserve the model protein's residue types. The conformation of the reference side chain is preserved as much as possible by aligning the dihedral angles in common with the residues after aligning of the backbone.

The coordinates for the loop or variable regions are generated by finding peptide segments in other proteins that fit properly into model's spatial environment. The peptide segments are found by the 'search loop' command by searching the Brookhaven protein database which meet a defined geometric criterion. The alpha-carbon distance matrix is used to search for the segment in which having the best fit of the alpha-carbon distance with the specific number of residues. Best fit is defined as the lowest root mean square distance value as shown in equation 2.1.

$$\left(\sum_{i=1}^{N}\frac{(x-x_0)^2+(y-y_0)^2+(z-z_0)^2}{N}\right)^{1/2} \qquad 2.1$$

A distance difference between the distance matrix ($d_{i,j}$) and the distance matrices of all possible loop candidates (c) calculated over all N preflex (pe) and postflex (po) positions is shown in equation 2.2.

$$D_c^2=\frac{2}{N(N+1)}\sum_{i\in pe,po}\sum_{j\in pe,po,j>1}\left(d_{i,j}-d^c_{r_c(i)r_c(j)}\right)^2, i,j\in pe,po \qquad 2.2$$

Ten best models are proposed for the loop regions where the loop 1 has the best fit while loop 10 has the poorest. A flex region is defined as the portion of the molecule which is not included in the search and its geometry was allowed to vary. The residues which are leading up and going away from the flex are defined as preflex and postflex residues, respectively. An appropriate loop is selected and automatically incorporated into the model. Figure 2.5 shows geometry definition for the search loops command.

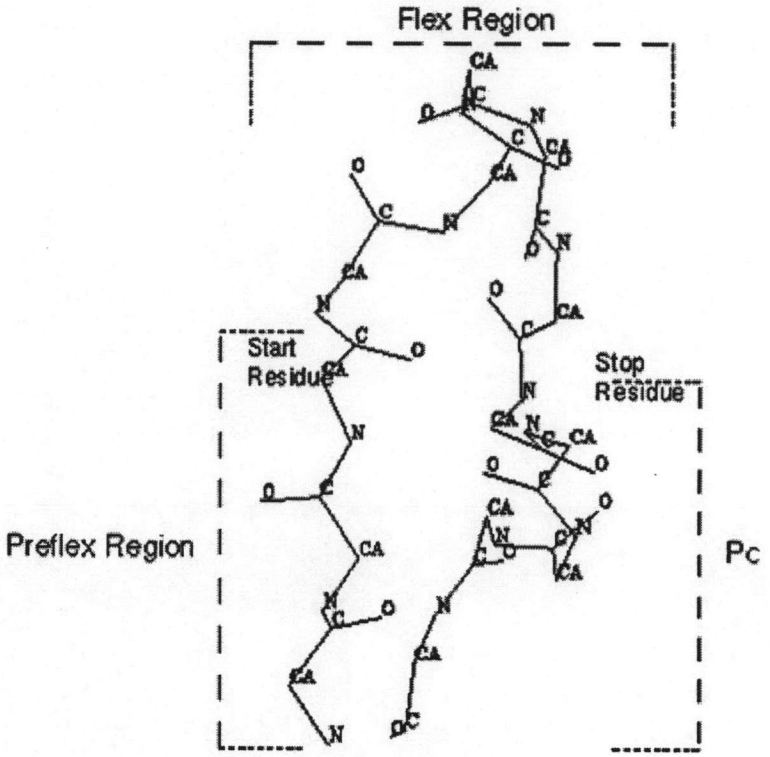

Figure 2.5 Geometry definition for the search loops command **(67)**.

2.1.2 Energy minimization

As for the model building, many artifacts were introduced into the model protein which includes substitution of large side chains, strained peptide bonds between segments taken from different reference proteins as well as the non-optimum conformation for the loop. All of these artifacts can be relaxed using energy minimization.

The energy minimization was subjected to identify the stable conformations. In the computer simulations, two first-order minimization algorithms are frequently used, steepest descents and conjugated gradient **(3)**.

2.1.2.1 Steepest descents method

In this method, the energy was minimized by repeating minimization along the direction of the force. The first derivative of the potential energy

surface was used with respect to the Cartesian coordinates. The method moves in the parallel direction to the net force by considering moving down the steepest slope of the interatomic forces on the potential energy surfaces. By adding an increment to the negative gradient of the potential energy or forces, the descent is accomplished (see Figure 2.6 a).

The method is commonly used in the initial step for relaxing the poorly refined structure either resolved from the crystallography or model building since it reasonably converges at the initial step and requires minimal computing time. However, the progress becomes slow when approaching the minimum. This leads to another minimization method known as conjugate gradient to be used.

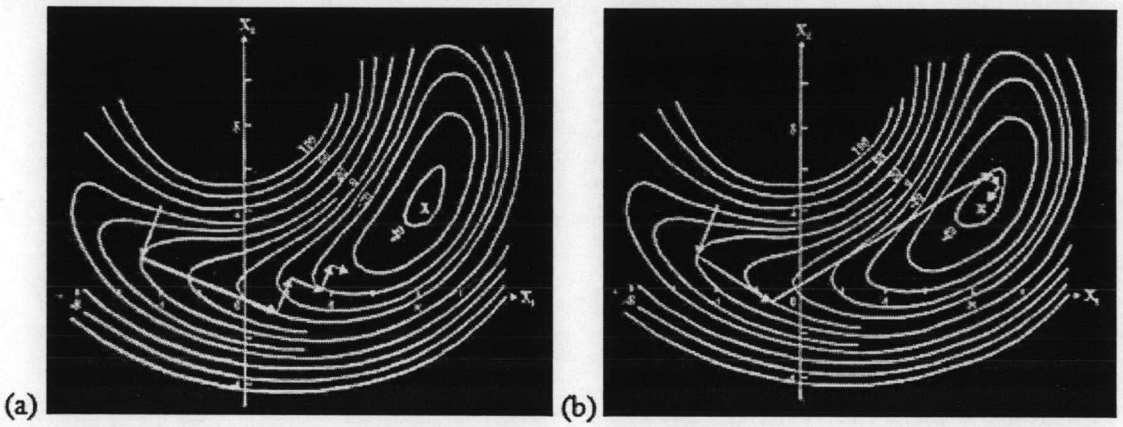

Figure 2.6 First-order minimization algorithm (a) steepest descents and (b) conjugate gradient methods **(69)**.

2.1.2.2 Conjugate gradient method

The method uses both current gradient and the previous search direction to drive the minimization. It is considered to select a successive search direction in order to eliminate repeated minimization along the same direction. This made the method faster than that of the steepest descent one. The conjugate gradient method is displayed in Figure 2.6 b.

2.2 Molecular dynamics (MD) simulation

Nowadays, computer plays critical important roles in the study of motion and dynamics properties of many systems. Molecular Dynamics (MD) simulation is one of the most useful techniques in computer modeling and has been applied to study dynamical behavior of various systems especially biomolecular system particular protein. For most proteins, the biological function includes an interaction with one or more small molecules (a ligand, substrate and inhibitors) or other macromolecules (protein, carbohydrate, lipid and nucleic acid). An investigation of the dynamics of the structural fluctuations and their relation to their reactivity and conformational changes has led to the understanding in details of the activity of such proteins and systems. The MD simulations have been instrumental in providing useful knowledge for the study in these fields. The dynamics obtained from MD studies are utilized to determine thermodynamics properties as well as information concerning their motions. One of the most special interests is to understand protein particularly enzyme stability and the thermodynamics of their interaction with drugs **(1-3)**.

2.2.1 Theory

MD simulation technique is based on the classical equation of motion, *i. e.* Newton's equation (equation 2.3). Basically, from the force acting on each atom, the acceleration of each atom in the system can then be determined. By integrating the equation of motion, a trajectory describing the positions, velocities and the accelerations over time can be obtained. The average values of many properties can then be determined from this trajectory.

$$F = ma \tag{2.3}$$

where F is the force exerted on particle, m is particle mass and a is the acceleration of the particle and equal to the second derivative of the position over time.

The force can be expressed as the gradient of the potential energy.

$$F = -\nabla V \tag{2.4}$$

where V is the potential energy of the system.

By combining equation (2.3) and (2.4), the Newton's equation of motion thus relate the derivative of the potential energy to the changes in position as a function of time, equation (2.5)

$$m\frac{d^2r}{dt^2} = -\frac{dV}{dr} \tag{2.5}$$

Since, the acceleration equals to the derivative of velocity over time and velocity equal to the derivative of position over time, the acceleration can then be related to the position, velocity and time as follow:

$$\begin{aligned} a &= \frac{dv}{dt} \\ v &= at + v_0 \\ v &= \frac{dx}{dt} \\ x &= vt + x_0 \\ x &= at^2 + v_0 \cdot t + x_0 \end{aligned} \tag{2.6}$$

where x is the position of the particle, x_0 and v_0 are initial position and velocity, respectively. To calculate the trajectory, the initial positions of atoms, initial distribution of velocities and the acceleration are needed. The acceleration can be determined from the gradient of the potential energy while the initial position can be obtained from experimentally resolved structures. The initial distribution of velocities is resolute from a Maxwell-Boltzmann distribution at a given temperature (equation 2.7).

$$\begin{aligned} p(v) &= \left(\frac{m}{2\pi k_B T}\right)^{1/2} \exp\left[-\frac{1}{2}\frac{mv^2}{k_B T}\right] \\ T &= \frac{1}{(3N)}\sum\frac{|P|}{2m} \end{aligned} \tag{2.7}$$

where N is the number of atoms in the system.

The integration algorithm approximated by a Taylor series expansion (equation 2.8) was introduced to solve the equation numerically as for the complicated system (function of atomic position) that does not have analytical solution.

$$\begin{aligned} x(t+\delta t) &= x(t)+v(t)\,\delta t+\frac{1}{2}a(t)\,\delta t^2+\ldots \\ v(t+\delta t) &= v(t)+a(t)\,\delta t+\frac{1}{2}b(t)\,\delta t^2+\ldots \\ a(t+\delta t) &= a(t)+b(t)\,\delta(t)+\ldots \end{aligned} \qquad (2.8)$$

There are many integration algorithms available but, here, we will show the algorithm which is used in the AMBER program, so called *leap frog* algorithm (equation 2.9). The velocities are first calculated from $t+1/2\delta t$ and then used to calculate the position x at time $t+\delta t$. The velocities at time t can then be approximated by equation 2.10.

$$\begin{aligned} x(t+\delta t) &= x(t)+v(t+\frac{1}{2}\delta t)\,\delta t \\ v(t+\frac{1}{2}\delta t) &= v(t-\frac{1}{2}\delta t)+a(t)\,\delta t \end{aligned} \qquad (2.9)$$

$$v(t) = \frac{1}{2}\left[v(t-\frac{1}{2}\delta t)+v(t+\frac{1}{2}\delta t)\right] \qquad (2.10)$$

The MD method is deterministic in which the state of the system at any time are predictable once positions and velocities of each atom are known. MD simulations are sometimes time consuming and computationally expensive. Nevertheless, the faster and cheaper of the computer today bring up the calculation to the nanosecond time scale.

The basic step in MD simulations was shown in Figure 2.7. The initial system, in which the coordinates can be normally obtained from X-ray or NMR data or built up using Molecular Modeling method, is minimized in order to get rid of bad atomic contacts due to the addition of some newly atoms and residues *i.e.* hydrogen atoms, amino acid residues as well as water molecules. The initial velocities of the system were

assigned to the system using Maxwell-Boltzmann distribution as described above. Prior to the production dynamics, the heating and equilibration dynamics were performed with the aim to decrease the probability that localized fluctuations in the energy such as hot spots in which will persist throughout the simulation. Following the initial Thermalization/equilibration, some more period of time for equilibrations were applied until the system is reached the equilibrium by adjusting the temperature and rescaling the velocities. The trajectory of the systems is calculated and used for analysis.

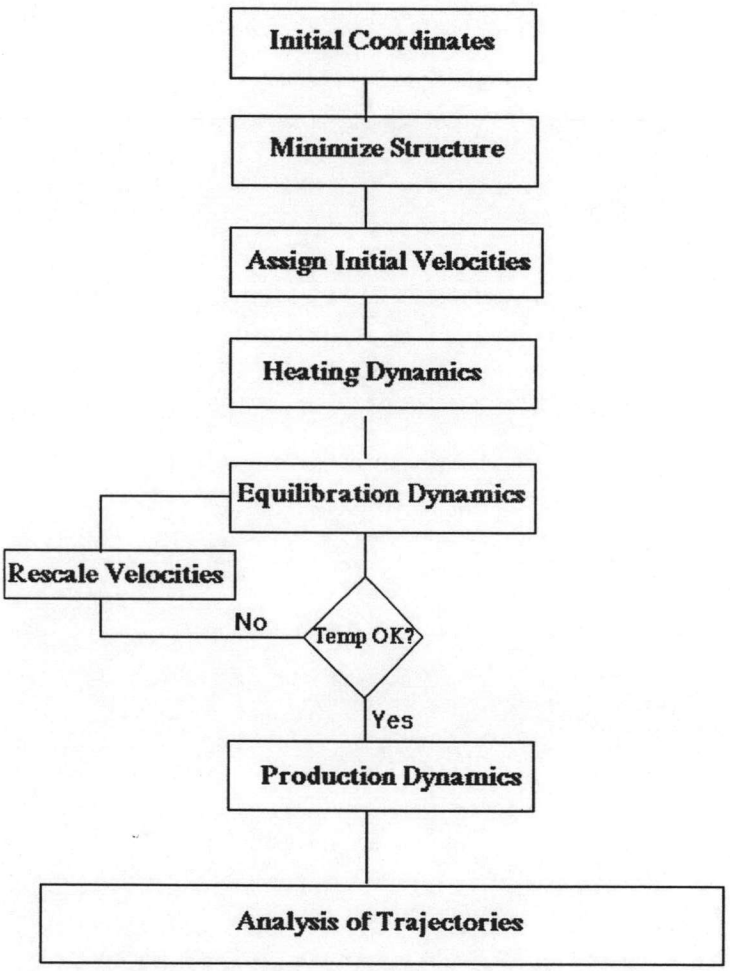

Figure 2.7 Basic steps on MD simulation study.

2.2.2 Potential function (force field)

Since the biological molecules are usually large, *i.e.*, involve many atoms, the study of the structure, function and dynamics relationship at the atomic level cannot

be accomplish at the quantum mechanics level **(1-3)**. The empirical potential energy function so called force field, which is much less computationally time consuming, is, in stead, used for such a large system. The function provides reasonable results compromise between the accuracy of the results and computationally efficiency.

Equation 2.11 displays the simplest functional form of the force field that represents the essential nature of the molecules in condensed phase. The potential energy, *V(R)* equation 2.11, is the summation of the bonded and non-bonded interactions. The bonded interactions come from the summation of the bond energy (first term), the angle energy (second term) and the dihedral angle energy (third term). While, the van der Waals interaction (fourth term) and the electrostatic interaction (last term) form the non-bonded interaction.

$$
\begin{aligned}
V(\mathrm{R}) \quad) = & \sum_{bonds} K_r \, (r - r_{eq})^2 \\
& + \sum_{angles} K_\theta \, (\theta - \theta_{eq})^2 \\
& + \sum_{dihedrals} \frac{V_n}{2} (1 + \cos[n\phi - \gamma]) \\
& + \sum_{i<j}^{atoms} \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{\ 6}} \\
& + \sum_{i<j}^{atoms} \frac{q_i q_j}{\varepsilon R_{ij}}
\end{aligned}
\tag{2.11}
$$

where K_r, K_θ are stretching and bending force constants, r_{eq} and θ_{eq} are equilibrium values for bond and angle, V_n is the rotational barrier height, n is the periodicity of rotation, ϕ is torsional angle, γ is the phase angle, q represents point charges of atom i and j, ε is dielectric constant and R_{ij} is distance between atom i and j.

2.2.3 Periodic boundary condition

The periodic boundary condition was introduced in the MD simulations for reducing the boundary effect at the edge of the simulation box where the molecules which stay close to the edge of the box receive the incomplete interaction (one side interaction) **(1-3)**. The periodicity was applied to the system of interest in which locating at the central cell. The interactions of the molecule nearby the edge of the box were then complete by involving the interaction from the next box. Furthermore, the number of molecules was kept constant by entering of the image of the leaving cell.

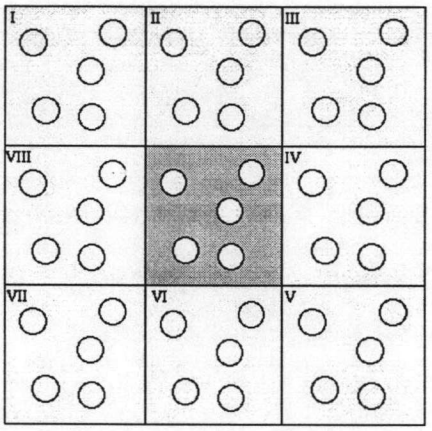

Figure 2.8 The periodic boundary condition in two dimensions.

2.2.4 Potential cut-off for non-bonded interaction

The calculation of non-bonded interactions is the most time-consuming part in the MD simulation since all pair interactions are calculated for every pair of atoms in the system. In order to reduce the calculation time, the total forces acting on a particle from neighboring particles are of the main contribution. The interactions are evaluated between each pair of particles with a distance less than a cut-off radius, r_{co}. This compromises between the correction and efficiency **(1-3)**.

2.2.5 Treatment of long range interaction

The Particle Mesh Ewald (PME) method was used to calculate the full electrostatic energy of a periodic box in a macroscopic lattice of repeating images **(70)**.

The method is based on the Ewald summation method **(71)** and particle mesh method **(72)**, which gives the exact result for the electrostatic energy of a periodic system containing an infinite replicated neutral box of charged particles. The method is usually used for the complex molecular system with periodic boundary condition.

Consider the Coulomb's law:

$$V_{r_{ij}} = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \tag{2.12}$$

where $\varepsilon = 4\pi\varepsilon_0$, the Ewald method splits the Coulomb potential which is slowly converged into three exponentially converged contributions as described in equation (2.13)

$$V_t = V_f + V_r + V_s \tag{2.13}$$

where V_t, V_f, V_r and V_s are potential for total, Fourier Space part, Real Space part and Self interaction, respectively. Each contributor is given as follows:

$$\begin{aligned} V_f &= \frac{1}{2v\varepsilon_0}\sum_{k\neq 0}^{\infty}\frac{\exp[-k^2/4\alpha^2]}{k^2}\left|\sum_{i=1}^{i=N}\exp[-ik\cdot r_i]\right|^2 \\ V_r &= \frac{1}{4\pi\varepsilon_0}\sum_{i<j}\frac{q_i q_j\, erfc\,(\alpha r_{ij})}{r_{ij}} \\ erfc(x) &= \frac{2}{\sqrt{\pi}}\times\int_x^{\infty} dt\exp[-t^2] \\ V_s &= -\frac{1}{4\pi\varepsilon_0}\sum_{i=1}^{i=N}\frac{q_i^2\alpha}{\sqrt{\pi}} \end{aligned} \tag{2.14}$$

Application of the PME method allows fast Fourier transform (FFT) to be applied, thus, the method is better.

2.3 Molecular docking

It is generally known that ligands, which are substrate or inhibitor (drug), usually bind to receptor (protein) at a cavity of the receptor called binding site. Understanding the

nature of binding as well as the interaction between proteins and their ligands are of important in many biological systems especially in the epidemic diseases. In order to understand such phenomenon, the complex structures of both molecules need to be known. The structure of the complexes between proteins and ligands can be determined by crystallographic techniques. However, many systems are difficult to solve. This made a molecular docking, a computational method, plays more crucial role in the structure-based drug designing field (Figure 2.9).

It is assumed that the geometry and the energetic factor are of important in the binding process. Considering the small molecules ligand with high degree of freedom, proteins are much bigger in term of number of atoms and sizes. Given the protein and ligand finding the position and configuration to bind, it is obviously a difficult problem to solve, i.e. high degree of freedom. Many approximations are thus introduced to make the computation feasible, for example, protein is considered to be rigid, only some parts of ligand can rotate etc **(73)**.

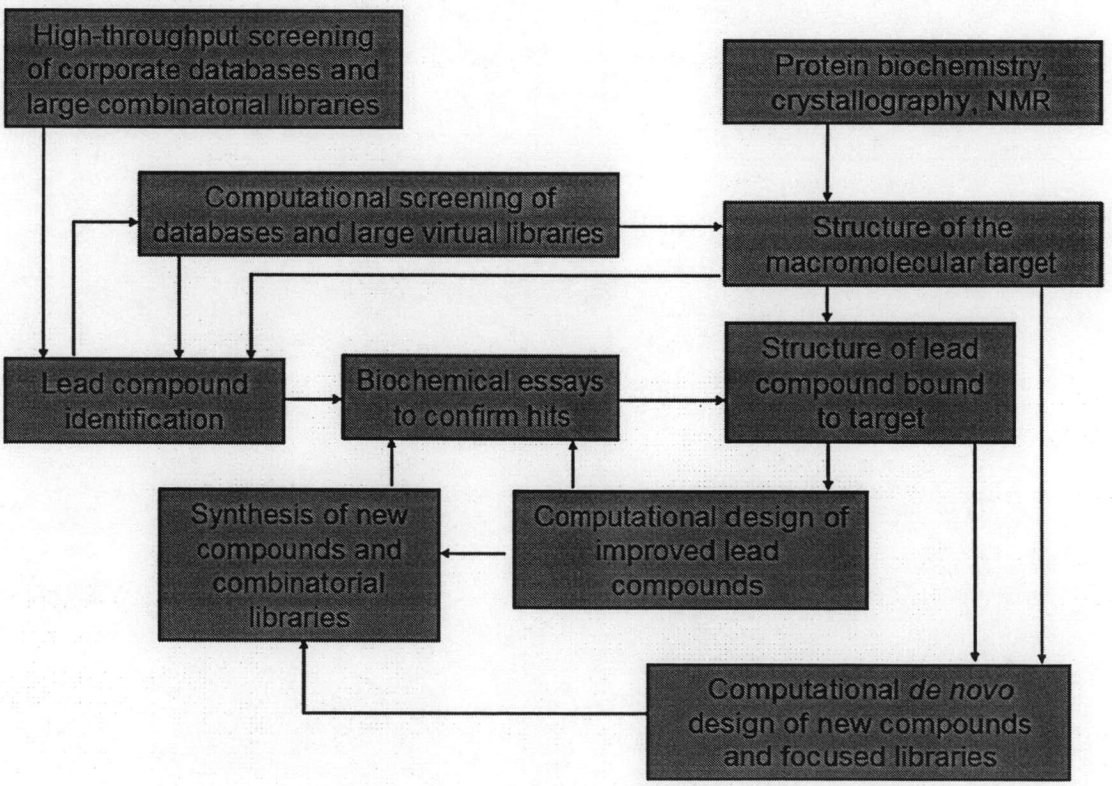

Figure 2.9 Schematic of the structure-based drug design process.

Hex is an interactive protein docking and molecular superposition program **(74)**. The basic approach of the program is to represent the steric shape, electrostatic potential and charge density of each protein as expansions of spherical polar Fourier basis functions **(75)**.

2.3.1 Gaussian density representation of protein shape

Gaussian density representation of protein shape **(76)** is used to enhance the original shape sampling algorithm **(77)**.

$$\rho_i(\underline{r}) = \alpha \exp\{-\beta(r / r_i)^2\} \qquad 2.15$$

where $\rho_i(\underline{r})$ is the density function for atom i, r_i is its van der Waals (VDW) radius, and α and β are adjustable parameters.

2.3.2 Fourier expansion and coordinate operations

A Fourier expansion to order N of property $A(\underline{r})$ in spherical polar coordinates $\underline{r} = (r, \theta, \phi)$ is described in equation 2.16 **(74)**.

$$A(\underline{r}) = \sum_{nlm}^{N} a_{nlm} R_{nl}(r) y_{lm}(\theta,\phi); \quad N \geq n > l \geq |m| \geq 0 \qquad 2.16$$

where a_{nlm} is an expansion coefficient, $R_{nl}(r)$ represents either an harmonic oscillator or a Coulomb-type radial function, and $y_{lm}(\theta,\phi)$ is a real spherical harmonic.

The rotational and translational coordinate operations on spherical polar Fourier expansions are represented as equation 2.17 and 2.18.

$$\hat{R}(\alpha,\beta,\gamma)A(\underline{r}) = \sum_{nlm}^{N} a'_{nlm} R_{nl}(r) y_{lm}(\theta,\phi) \qquad 2.17$$

$$\hat{T}_z(R)A(\underline{r}) = \sum_{nlm}^{N} a''_{nlm} R_{nl}(r) y_{lm}(\theta,\phi) \qquad 2.18$$

where the rotated and translated expansion coefficients are respectively given by equation 2.19 and 2.20.

$$a'_{nlm} = \sum_{m'=-l}^{l} R^{(l)}_{mm'}(\alpha,\beta,\gamma) a_{nlm'} \qquad 2.19$$

$$a''_{nlm} = \sum_{n'l'}^{N} T^{(|m|)}_{nl,n'l'}(R) a_{n'l'm} \qquad 2.20$$

In a rigid body docking search, it is convenient to represent steric and electrostatic complementarily as overlap integrals between corresponding pairs of 3D functions. In case both molecules are initially located at the origin, the correlation S_{AB} between any pair of functions $A(\underline{r})$ and $B(\underline{r})$ for molecules A and B, respectively, is calculated using equation 2.21.

$$S_{AB}(R,\beta_1,\gamma_1,\alpha_2,\beta_2,\gamma_2) = \int [\hat{T}_z(-R)\hat{R}(0,\beta_1,\gamma_1)A(\underline{r})] \times [\hat{R}(\alpha_2,\beta_2,\gamma_2)B(\underline{r})]dV \qquad 2.21$$