Chapter 4

In-silico mutations, molecular dynamics simulations and docking studies on WCI mutants: Principles and Methods

Protein structures are primarily stabilized by a large number of weak non-bonded interactions. The fluctuations of the atomic positions in the protein control the nature of these interactions. Variations in the properties of the amino acids such as polarity, hydrophobicity or ability to form hydrogen bonds modulate the stability of the different regions of the proteins and eventually lead to the differences in the fluctuation amplitudes. One of the approaches to study the motional properties of biological macromolecules, like protein or DNA is provided by the molecular dynamics (MD) simulations, which essentially calculates the fluctuations in the relative positions of the atoms as a function of time (Karplus & McCammon, 1983; Petsko & Ringe, 1984). In other words, MD simulations provide individual particle motions as a function of time so that they can be probed far more easily than experiments to answer detailed questions about the properties of the systems.

Protein-protein interactions play a central role in various aspects of the structural and functional organization of the cell. Computation techniques like protein-protein docking can elucidate the details of these interactions at the atomic level. MD simulations, when aided with docking

studies, can measure the effect of atomic fluctuations of the proteins in their recognition processes and interface interactions.

4.1. Objective of *in-silico* mutations and MD simulation in WCI

In the previous chapters, I discussed about the X-ray structural studies on the two mutants of WCI, N14K and N14D. The crystal structure analysis of these two mutants (N14K at 2.05Å and N14D at 1.9 Å resolutions) established the necessity of replacing this residue (Asn14) with few other residues of different size and electrostatic nature to understand its role in more details. It is known that well converged MD simulations include entropic effects and can be used as a powerful tool for the conformational search in the loop regions while restraining the non-loop residues close to the initial model (Hornak and Simmerling, 2003). In this work, we have replaced Asn14 with some other residues *in-silico* and MD simulations have been carried out on the reactive site loop region of native WCI and its mutants, to explore the role of Asn14 and to understand the dynamic nature of the reactive site loop of the protein. The residues having long/bulky side chains were excluded as in case of N14K, we observed that the side-chain of Lys14 was not accommodated in the loop cavity as such and folded back (Chapter 3) with an unusual conformation.

To start with, Asn14 was replaced by Gly (N14G), where the side chain is fully truncated, followed by Ala (N14A), having just a methyl group as the side chain. Subsequently the mutations with hydrophobic residues like Val (N14V) and Leu (N14L) were also carried out which eventually prompted us to replace Asn14 with the residues having small but not-so-polar side chains like Thr (N14T) and Ser (N14S).

The simulated structures, obtained in each case, were then analyzed in terms of hydrogen bonding interactions, dihedral angles and water-orientation in the loop regions followed by docking in the active site cleft of cognate enzyme α-chymotrypsin using the program MULTIDOCK (Jackson *et al.*, 1998). These studies helped us to quantify the deviations in the canonical conformation in terms of interaction energy and the occurrence of conserved hydrogen bonds as seen in case of serine protease-protease inhibitor complexes.

4.2. Principles of MD simulation

MD simulations begin with the knowledge of energy (of the system) as a function of atomic coordinates. The potential energy surface determines the relative stabilities of different possible stable or metastable structures. The force acting on the atoms of the system, which are related to the first derivatives of potentials with respect to the atom positions, can be used to characterize the dynamic behavior of the system by solving Newton's equation of motion for the atoms as a function of time (Karplus and Petsko, 1990).

For a simple homogeneous system, such as a box of water molecules with periodic boundary conditions, average structural and dynamic properties can be determined in simulations only of a few picoseconds but inhomogeneous systems like proteins require considerably longer simulations. Modern computers allow simulations of upto nanoseconds, long enough to completely characterize the atomic fluctuations of the proteins.

The energy functions used for proteins are generally composed of bonding terms representing bond lengths, bond angles, torsion angles, and non-bonding terms consisting of van-der-Waals interactions and electrostatic contributions. One widely used energy expression Brookes *et al.*, 1983) is:

$$E(R) = 1/2 \sum K_b (b-b_0)^2 + 1/2 \sum K_\Theta (\Theta-\Theta_0)^2 + 1/2 \sum K\phi [1+ \cos (n\phi - \delta)]$$

$$+ \sum (A/r^{12} - B/r^6 + q_1q_2/Dr) \qquad \text{.....Equation (1)}$$

The energy E, is a function of the Cartisian coordinate set, R, specifying the positions of atoms. The first term in the above equation represents instantaneous displacements from the ideal bond length, b_0, by a Hooke's law (harmonic) potential. Such a harmonic potential is the first approximation to the energy of a bond as a function of its length. The bond force constant K_b determines the flexibility of the bond and can be evaluated from infrared stretching frequencies or quantum mechanical calculations. The energy associated with alteration of bond angles, given by the second term in the equation, is also represented by a harmonic potential. For rotation about bonds, torsion angle potential functions (third term in equation 1) are used. The final term in equation (1) represents the contribution of non-bonded interactions and has three parts: a repulsive term preventing atoms from interpenetrating at very short distances; an attractive term accounting for the London dispersion forces between atoms; and an electrostatic term that is attractive or repulsive depending on whether the charges q_1 and q_2 are of opposite or the same sign. The first two non-bonded terms combine to give the familiar Lennard-Jones 6-12 potential, which has a minimum at an inter-atomic separation equal to the sum of the van der Waals radii of the atoms; parameters A and B depend on the atoms involved and have been determined by a variety of methods, including non-bonding distance in crystals and gas-phase scattering measurements. Electrostatic interactions between pairs of atoms are represented by a Coulomb potential with D the effective dielectric function for the medium and r the distance between the two charges. Use of atomic partial charges avoids the need for a separate term to represent the hydrogen bond interactions.

4.3. *In-silico* mutations and MD simulations: procedure

The Molecular dynamics simulations were performed on WCI and its mutants in a Silicon Graphics O2 Workstation, using **DISCOVER-III** module of **Insight-II** from Molecular Simulation Inc (MSI) (San Diego, CA). For all the computations, described in this study, the X-ray coordinates of WCI (PDB code:1EYL), excluding the water molecules and sulphate ions were used as the starting model. Asn14 of WCI was mutated *in-silico* to Gly, Ala, Ser, Thr, Val and Leu using the Insight-II **Biopolymer** module and the lowest energy rotamer for the mutated residues were chosen, whenever necessary.

To start with, hydrogen atoms were generated at pH 7.0 (according to idealized bond lengths and valance angles) using **BUILDER** module of **InsightII**. Disulfide bonds were also generated using the same module. Potentials were assigned using consistent valance forcefield (CVFF) with a distance independent dielectric constant of 4.0. No cross-terms were used in the energy expressions. For the treatment of non-bonded interactions, atom based method was used where non-bonded interactions were evaluated with a cutoff distance of 9.5 Å. The hydrogen atom positions were energy minimized using the steepest descent algorithm (down to a gradient of <100 kcal mol^{-1} $Å^{-1}$) followed by conjugate gradient minimization (down to <10 kcal mol^{-1} $Å^{-1}$ and then 0.001 kcal mol^{-1} $Å^{-1}$) keeping the heavy atoms of the proteins fixed. Next the subsets were defined and the protein was subjected to further minimization and simulations.

4.3.1. Subset define

For molecular dynamics simulations, when applied to the large systems like proteins, the required computer time becomes almost prohibitive. One strategy commonly used for

simplifying the treatment of large systems is to restrict the computation to a subset of atoms around the site of interest (Brookes *et al.*, 1988; Karplus and Petsko, 1990). In MD simulation studies done here, a subset, SUB1 was defined in each case, which initially included the reactive site loop residues Arg59-Val75 and the 14^{th} residue. But the three dimensional structure of WCI reveals that the residues of extended reactive site loop region interact with few other N-terminal residues. Therefore to get a real insight into the interaction scenario at the reactive site loop region, the N-terminal residues Asp1-Gly16 were also included in SUB1 (Fig. 4.1; coloured green). The rest part of the protein was defined as the subset, SUB2 (Fig. 4.1; coloured brown) which was kept fixed during subsequent minimizations and MD simulations.

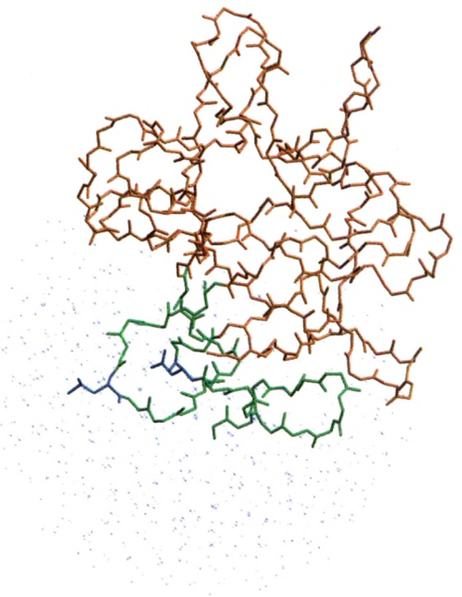

Figure 4.1. Subset SUB1 is defined around the reactive site loop region (green) and the rest part of the protein is defined as SUB2. SUB1 is solvated and the water molecules are shown in sky-blue. Asn14 and Leu65 (P1) are coloured in blue.

4.3.2. Solvation of the subset

It is known that proteins and nucleic acids have evolved to function in an aqueous environment; large numbers of water molecules are tightly bound to their surfaces, forming a first hydration shell which is an integral part of the structure. Crystal structures of N14K and N14D revealed the role of water molecules in stabilizing the reactive site loop conformation (Chapter 3, Fig. 3.3). Hence in our present simulation study, it is also necessary to understand the structural and dynamic properties of solvent-protein interaction. Moreover, simulations *in vacuo* suffer from serious artifacts, such as an excessive deviation from the native conformation. Therefore, we decided to perform the simulations on the solvated protein molecules. The **SOAK** utility of **DISCOVER-3** module was used for this purpose. A water sphere of 21Å radius was generated considering C_α atom of Ser66 as the centre of the sphere. The water molecules thus generated (Fig. 4.1; coloured blue) were energy minimized using the steepest descent algorithm (down to a gradient of <100 kcal mol^{-1} $Å^{-1}$) followed by conjugate gradient minimization (down to <10 kcal mol^{-1} $Å^{-1}$ and then 0.001 kcal mol^{-1} $Å^{-1}$) successively. Finally, the energy minimized water molecules were simulated for 30ps keeping the protein molecules fixed.

4.3.3. Energy minimization and MD simulation

The energy minimization followed by MD simulations, were then carried out on the solvated SUB1, keeping the rest of the protein molecule fixed. Energy minimization of the solvated subset was performed in each case by the steepest descent followed by conjugate gradient algorithm, as described in previous section. This minimized assembly of SUB1 and

solvent was then used as a starting point for NVT (constant volume and temperature) molecular dynamics to generate possible stable conformations. The simulations were carried out using the Verlet velocity algorithm (Swope and Anderson, 1982).

The usual disadvantage in MD simulations, done here, is that the computational system generally does not permit simulations long enough. In order to partially overcome this disadvantage, high temperature MD simulations were carried out. The idea of performing high temperature dynamics was to bring the *in-silico* mutated systems in a state of energy minima where it would be free from the biasness of starting (native WCI) model, before doing the production run for analysis. Hence the kinetic energy for MD simulations was provided initially by a thermal bath at a constant temperature of 300K and then the temperature was elevated upto 500K. In each case, a total simulation run of 850ps was performed with 1fs time step. Simulation parameters are summarized in Table 4.1.

Choosing of 850ps simulation was not arbitrary. At least in two cases (WCI and N14L) 1 ns dynamics were done and analyses showed that the extra run after 850ps did not add any further information. As we were dealing with a number of mutants, we wanted to optimize the CPU time and settled for 850ps MD run (100ps production run) where the initial run of 250ps at 300K was followed by a 400ps run at 500K and again a 200ps run at 300K. After a 200ps of thermalization at 500K, when the system reached a state with minimum fluctuations of potential energy among the individual time steps, a trajectory of 200ps at 500K followed by another 200ps at room temperature have been performed. The potential energy vs. time plot was carefully monitored after each 100ps run and the conformations were saved after every 0.2ps. The last 100ps dynamics run at 300K has been considered as production run for analysis.

MD simulation protocol, in brief:

Generation of Hydrogen atoms → Assignment of force-field

→ Minimization of hydrogen atoms keeping the heavy atoms fixed

→ Subset define (SUB1) around the reactive site loop

→ Rest part is defined as SUB2

→ Generation of water molecules and solvation of the subset, SUB1

→ Minimization and 30 ps simulation of water molecules

→ Minimization of (SUB1+ water) asssembly

→ MD simulation on (SUB1+ water) assembly

Table 4.1. Summery of MD simulation parameters

Force field	CVFF
Dielectric constant (Distance independent)	4
Summation method	Atom based
Spline width	1.0 Å
Buffer width	0.5
Ensamble	NVT
Cutoff distance	9.5 Å
Total simulation time (each)	850 ps
Time step	1 fs
Movement limit in minimization	0.2
Temperature control method	Velocity scale

4.4. Principles of protein docking

All biological processes in living organisms involve molecular recognition between interacting partners (protein-protein, protein-DNA, protein-ligand etc). Molecular recognition is mediated by electrostatic and surface complementarity between associating biomolecules. Understanding the principles underlying molecular recognition is of great importance in drug designing and specific inhibition of selected proteins in a pathway. Most of these principles in terms of complementarity and molecular interactions, stabilizing heterodimeric protein structures, have been incorporated into computational docking algorithms to predict a priori the affinity and geometry of association of proteins with their corresponding ligands and inhibitors.

Though the recent years have seen a dramatic increase in the number of three dimensional structures of proteins, the structure determination of protein-protein complexes is still considered a difficult problem in many cases. The non-availibility of complex structures necessitates the development of modeling and docking techniques (Cherfils & Jenin, 1993; Janin, 1995; Shoichet & Kuntz, 1996) in which the interactions between the members of a molecular complex can be examined. Methods for computer aided molecular docking include a reasonably accurate estimation of energy and also deal with the combinatorial conformational complexity incurred by molecular flexibility of the docking partners. Protease-inhibitor systems appear to be the most amenable to prediction because of the rigidity of the protein backbone despite molecular association, even though the side chain can undergo considerable conformational changes (Betts & Sternberg, 1999). Predictive docking has been successfully used in case of proteases and their inhibitors for understanding the interactions at their recognition site.

Programs available for protein-protein docking can be divided into two different classes: 'rigid-body docking (Shoichet & Kuntz, 1991; Cherfils, 1991; Jiang & Kim, 1991; Bacon & Moult, 1992; Walls & Sternberg, 1992; Katchalski-Katzir *et al.*, 1992; Fisher *et al.*, 1995; Gabb *et al.*, 1997) and 'flexible' docking (Jackson *et al.*, 1998; Weng *et al.*, 1996; Cumming *et al.*, 1995). In case of rigid body docking, the coordinates of the proteinases and their inhibitors are constrained whereas in flexible docking the conformational changes of the residues constituting the recognition sites are considered. Inclusion of molecular flexibility to global optimization is relatively uncommon because it needs enhanced computational time.

In the present study, the associations of WCI and mutants with the cognate protease α-chymotrypsin are elucidated by means of docking using the program MULTIDOCK.

4.5. Docking: procedure

For WCI and the mutants, twenty snapshots with an interval of 5ps were collected in each case from the last 100ps production run and used for the docking studies with Chymotrypsin. Two different serine protease inhibitor complexes, STI-PPT (pdb code: 1AVW, Song and Suh, 1998) and OVO-CHY (pdb code: 1CHO, Fujinaga *et al.*, 1987) were used as initial templates for the docking experiments. In case of STI-PPT as template, the chymotrypsin molecule was first superposed on PPT and then the backbone of the (P3-P2′) region of reactive site loop of each snapshot was superposed with the corresponding part of STI to bring it to the peptide-binding groove of the enzyme. For OVO-CHY as template, only the reactive site loop (P3-P2′) of each snapshot was superposed on the corresponding part of OVO and the complexed chymotrypsin coordinate of OVO-CHY was used without any modification. The

program MULTIDOCK was used next to refine the interface of initial docked complex by rigid body minimization for better evaluation of the possible enzyme-inhibitor interactions.

4.5.1. Principles of MULTIDOCK

The program MULTIDOCK provides a method for refining the interface between two proteins (at the atomic level) of an initially docked complex, generated by a docking algorithm or manual docking procedures. This program models the effects of side-chain conformational change and rigid body movement of the interacting proteins during refinement.

The parameters used to define protein-protein interface and the non-bonded interactions in mean filed optimization and rigid-body minimization are as follows:

4.5.1.1. Definition of protein receptor and ligand in MULTIDOCK

In case of protein-protein inhibitor interactions, the PDB files sequentially consist of protein (receptor/enzyme) chain followed by the protein inhibitor (ligand) chain. The receptor/enzyme and inhibitor/ligand are defined in the *control_parameter* file by using the keywords ***imobile_mol*** (for receptor protein which remains fixed) and ***mobile_mol*** (for ligand/inhibitor protein which remains mobile). In this study, the transformed (superposed) coordinates of α-Chymotrypsin and mutant inhibitors were declared as ***imobile_mol*** and ***mobile_mol*** respectively.

4.5.1.2. Parameters related to protein-protein interface

The intermolecular interface (of interest) is described by a region with side chain mobility where multiple copies of side chain representations are modeled. This is surrounded by a fixed (buffer) region with existing side chain conformation. These two inclusion regions are defined in terms of the distance of C_β atom of a given residue (C_α for Glycine) from the C_β of any residue on the interacting protein. This is done by using the keywords: ***cut_iface*** (the distance cut-off in Å for inclusion in the region of the multiple copy side chain rotamer representation) and ***cut_jface*** (the distance cut-off in Å for inclusion as a fixed side chain rotamer representation).

4.5.1.3. Parameters that control non-bonded interactions

The control of non bonded interactions (which are identical in the mean field optimization and energy minimization steps) are defined using the following key words:

cut_nbond is the atom-atom non-bonded cut-off distance (in Å) for the calculations of residue-residue interaction energy.

d is the effective dielectric constant, ε, (no units) in V = 332.0*qi*qj/rij*ε

eatmax is the maximum van der Waals atom-atom interaction energy (in kcal/mol).

cut_xx is the distance of closest approach (in Å) for heavy atom contacts necessary to evaluate the electrostatic interaction energy.

cut_xh is the distance of the closest approach for heavy-hydrogen atom contacts to evaluate the electrostatic interaction energy.

cut_hh is the distance of the closest approach for hydrogen-hydrogen atom contacts to evaluate the electrostatic interaction energy.

4.5.1.4. Parameters specific to mean field optimization

Following are the parameters necessary for mean field optimizaion:

cut_res_nb is the residue-residue non-bonded cut-off distance (in Å) between adjacent Cβ atoms for the inclusion in the side chain residue-residue non-bonded pair list.

temp is the temperature parameter (in Kelvin) that determines the value of RT used in ensamble optimization.

lamda is the value for memory of previous probability matrix. The value of lamda ranges from 0.0 to 1.0. The smaller the value of lamda, the smaller the memory of previous probability matrix.

rmsmax determines the convergence criteria in terms of the r.m.s. change in the probability matrix. When rmsMAT is below rmsmax the probability matrix is deemed.

emax determines the convergence criteria in terms of change in energy (in kcal/mol).

4.5.1.5. Rigid-body energy minimization

Rigid body energy minimization is performed to relax the protein-protein interface. In our study, the maximum rotational (***thetamax***) and translational (***stepmax***) step sizes for rigid body minimization were 0.3° and 0.3 Å respectively. During this process, the parameter ***cut_lface*** maintained the residue-residue distance cut-off across the interface for the inclusion of the residues in the calculation of non-bonded interaction between the two rigid molecules.

Minimization continues until the energy of the system decreased by less than 10^{-6} kcal/mol for any given step. The resultant rotation and translation vectors were applied for updating the docked co-ordinates. The whole process of minimization was repeated until convergence of interface interaction energy was achieved.

Values of the important parameters of MULTIDOCK, used in the present study are summarized in Table 4.2.

Table 4.2. Summary of the parameters used in MULTIDOCK

cut_iface	12.0
cut_jface	24.0
Temperature	298.0K
cut_atom_nb	10.0 Å
cut_res_nb	15.0 Å
Dielectric	4
eatmax	2 kcal/mol
cut_xx	3.0 Å
cut_xh	2.0 Å
cut_hh	1.0 Å
Lamda	0.5
emax	0.2 kcal/mol
cut_lface	15.0 Å
thetamax	0.3°
stepmax	0.3 Å