APPENDIX 1

X-RAY CRYSTALLOGRAPHY

X-ray crystallography is a method of determining the arrangement of atoms within a crystal, in which a beam of X-rays strike a crystal and diffracts into many locations. From the angles and intensities of these diffracted beams, a three dimensional picture of the densities of electrons in the crystal can be generated. From this electron density, the mean positions of the atoms in the crystal can be determined. This is the only technique available for the determination of structure of biological macromolecules.

X-rays are generated when electrons collide with the atoms of a metal target such as copper. The electrons are liberated from a heated filament and accelerated by high voltage towards the metal target. X-rays are a form of electromagnetic radiation, where wavelength is about 0.1–10 nm. X-rays have been chosen to study the three dimensional structures of proteins and other macromolecules as their wavelength is on the same order of magnitude as the bond length of the atoms within molecules. The bond length between atoms within a molecule is about 0.15 nm or 1.5 Å. So, X-rays can be utilized to visualize the geometry and structure of molecules. Individual atoms in a molecule can diffract X-rays; however, they are weak scatterers of X-rays. So diffraction might be too weak to be detected by any instrument. However, this problem can be overcome by analyzing a diffraction pattern of a crystal rather than individual molecules. This is because a crystal is composed of a number of repeating patterns (unit cells) in a regular and ordered manner. Each molecule within the crystal therefore diffracts equally, and thus strong diffracted X-ray beams can be measured (Rhodes, 2000).
A.I.1. Bragg’s Law

W. L. Bragg visualized the scattering X-rays from a crystal by considering that the diffracted beams were reflected by planes passing through points of a crystal lattice. The diffracted X-rays are scattered by the crystal at a certain angle of reflection (θ). This reflection is analogous to that from a mirror, for which the angle of incident X-ray beam is equal to the angle of diffracted X-ray beam. The incident and the diffracted X-rays are in the same plane and the X-rays of wavelength (λ) are normal to a set of diffracting planes. The constructive interference between X-rays scattered from successive planes in the crystal will only take place if the path difference (2d) between the X-rays is equivalent to an integral number of wavelength λ. This is the Bragg’s equation.

\[ n \lambda = 2d \sin \theta \]

Figure A.I.1. The geometry of diffraction and its relationship to Bragg’s Law.

In Bragg's law, if the wavelength and the diffraction angle of a reflection are known, the perpendicular distance between the lattice planes in the crystal (interplanar spacing, d) can be easily calculated. As the angle increases, d
must become smaller for the path length to remain equal to one wavelength. The equation can be rearranged as

\[ d = \frac{n\lambda}{2 \sin \theta} \]

The minimum interplanar spacing \((d_m)\), where \(d_m = \frac{1}{2} (\sin \theta_{\text{max}})\), is also interpreted as the resolution of an electron density map. Since the maximum possible value of \(\sin \theta\) is 1, so the smaller the \(d_m\) value, the higher the resolution will be of the X-ray diffraction pattern. For instance, if the radiation used for X-ray generation has a wavelength of 1.54 Å, then the maximum resolution that can be observed with this radiation would be 0.77 Å (Blundell and Johnson, 1976; Glusker, Lewis and Rossi, 1994). Most proteins do not diffract better than 1.5 Å (Glusker, Lewis and Rossi, 1994). If a protein is diffracted to a high resolution level (above 2 Å), most of the fine structural features can be resolved.

**A.I.2. Asymmetric Unit, Space Group, Unit Cell and Bravais Lattices**

Crystals can be characterized by three elements to precisely define the arrangement, coordination, and periodicity of the fundamental unit of which they are composed. These three elements are symmetry properties, repetitive features and distribution of the atoms in the repeating unit. Protein molecules are inherently asymmetric. The asymmetric unit is the smallest component in the crystal. The asymmetric unit may consist of one molecule, part of a molecule or several molecules not related by symmetry. If only one molecule occupies a unit cell, then the cell itself is chiral and has no symmetry elements at all. This object is termed as the asymmetric unit because no part of it is systematically related to any other by crystallographic properties. That means it has no symmetry elements such as rotation axis or mirror plane. In most cases, the unit cell contains more than one identical molecule or oligomeric complexes (dimmer, trimer,
tetramer etc.) in an arrangement that produces symmetry elements. So, the largest aggregate of molecule(s) in a cell that possesses no symmetry element but can be juxtaposed on other identical entities by symmetry operation is called the asymmetric unit (Rhodes, 2000). A set of symmetry operations includes rotation, reflection, inversion, rotatory inversion, screw axes, glide plane, and translation. These operations can be applied to an asymmetric unit. Combination of all these elements in all possible ways generates a total of 230 unique, three-dimensional space groups of symmetry operation.

The unit cell is the basic building block of the crystal and is repeated infinitely in three dimensions. The directions of constructive interference depend only on the size and shape of the unit cell. The dimensions of a unit cell are designated by six parameters: the length of 3 unique edges (a, b, c) which run along x, y and z coordinates respectively, and three unique angles (α, β, γ) as indicated in FigureA.1.2.
In virtually all cases, a crystallographer is concerned only with the content of the individual unit cell and the coordinate of the atoms within the unit cell. There are 14 allowable unit cell types classified as Bravais lattice to distinguish their characteristics. The Bravais lattices themselves can be divided into five types of lattices, which are primitive (P), centered (C), body-centered (I), face-centered (F) and rhombohedral (R) (McPherson, 2003; Stout and Jensen, 1989). Any crystal can be regarded as being established by consecutively translational repetition of the unit cell and its content along a, b, c by distance |a|, |b|, |c| respectively, until a continuous three–dimensional array of repeated unit cells in a regular manner has been created (Glusker, Lewis and Rossi, 1994; Glusker and Trueblood, 1985; Rhodes, 2000). The simple symmetry operations and elements needed to describe unit cell symmetry are translation, rotation and reflection. The
symmetry of a unit cell is described in 230 space groups (like P212121). The space group is a group of symmetry operations consistent with an infinitely extended, regularly repeating pattern. Protein molecules are asymmetric since all amino acids except glycine have chirality. However, the D form of amino acids does not exist in proteins and only the L form does. Thus, there are less symmetry elements (mirror planes, inversion centers and glide planes) involved within the unit cell, and less space groups can be used to designate the protein. This limits the possible space groups to 65 out of the 230 mathematically possible ones (McRee, 1999). There are seven crystal systems used to classify the symmetry of the crystal, this corresponds to the seven fundamental shapes for unit cells, consistent with the 14 Bravais lattices as displayed in the table below (Glusker, Lewis and Rossi, 1994; Stout and Jensen, 1989).

<table>
<thead>
<tr>
<th>Crystal System</th>
<th>Bravais Lattices</th>
<th>Lattice</th>
<th>Angle</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Triclinic</td>
<td>P</td>
<td>a≠b≠c</td>
<td>α≠β≠γ</td>
</tr>
<tr>
<td>2 Monoclinic</td>
<td>P, C</td>
<td>a≠b≠c</td>
<td>α=γ=90°, β</td>
</tr>
<tr>
<td>3 Orthorhombic</td>
<td>P, C, I, F</td>
<td>a≠b≠c</td>
<td>α=β=γ=90°</td>
</tr>
<tr>
<td>4 Tetragonal</td>
<td>P, I</td>
<td>a=b≠c</td>
<td>α=β=γ=90°</td>
</tr>
<tr>
<td>5a Trigonal</td>
<td>P</td>
<td>a=b≠c</td>
<td>α=β=90°, γ=120°</td>
</tr>
<tr>
<td>5b Rhombohedral</td>
<td>R</td>
<td>a=b=c</td>
<td>α=β=γ=120°, 90°</td>
</tr>
<tr>
<td>6 Hexagonal</td>
<td>P</td>
<td>a≠b≠c</td>
<td>α=β=90°, γ=120°</td>
</tr>
<tr>
<td>7 Cubic</td>
<td>P, I, F</td>
<td>a=b=c</td>
<td>α=β=γ=90°</td>
</tr>
</tbody>
</table>

Table A.I.1. The seven crystal systems
A.I.3. Data Collection

For data collection purposes, the most important factor required for evaluating data quality is the completeness of the X-ray data including all the indices and their associated intensities, with their standard uncertainties (Dauter, 1999). Two factors that influence the data completeness are the geometric and informative content. The geometric factor, arising from the symmetry of crystal lattice and the detector setup, is a quantitative factor related to a number of variables including the approach of angular rotating method, the selection of the total rotation range appropriate for the crystal symmetry, crystal-to-detector distance, crystal mosaicity and beam divergence. The informative factor includes the quality of the data, the dynamic range of detector and the R-factor. The longer the exposure time, the greater the intensities and the signal-to-noise ratio, and the better the data quality obtained (Dauter, 1997 & Dauter, 1999). The X-ray data quality for macromolecular crystallography is assessed by a global indicator, the merging R-factor (R-merge) or symmetry R-factor (R-sym). The merging R-factor is defined by the following equation (Blundell and Johnson, 1976):

\[
R\text{-merge} (I) = \frac{\sum_{hkl} \sum_i i I_i (hkl) - \{I(hkl)\}}{\sum_{hkl} + \sum_i i I_i (hkl)}
\]

Where \(I_i(hkl)\) is all observed intensities and - \(I (hkl)\) is the average value of all observed intensities. The R-merge value will be between 20 – 40% while the signal/noise \((I/\sigma(I))\) falls around 1.0 to 2.0 (Dauter, 1999). The quantity of merging R-factor is almost universally used for evaluating X-ray diffraction data.
Protein crystals diffract X-rays much less than do the small molecule crystals; hence the diffraction intensity is weak. Thus protein data collection requires high intensity X-ray source and high sensitivity area detector. The image plate (IP) is considered to be the most suitable detector for acquiring protein data with a home source, as well as at synchrotron radiation facility. IP is a very sensitive detector, with wider dynamic range; it has high spatial resolution and high count rate capacity, which are the fundamental requirements of an X-ray area detector (Amemiya, 1997). The IP has enabled protein crystallographer to obtain very accurate data sets with reduced X-ray dosage and exposure time. The plate can be erased by exposure to intense white radiation and can be used repeatedly. The IP has a radius of 350 mm mounted on a solid base and a computer can control its movements through a controller box. Each collected image has a size of 2 MB, which can be stored in the computer. The processes of exposure, data collection, readout and storage of data are carried out automatically. The X-ray beam with wavelength, 1.54 Å CuKα is generated from rotating anode. X-ray generator usually operates at 50 kv and 100 mA. Now it is routine to collect the macromolecular data at cryogenic temperatures. This technique of flash cooling protein crystals at liquid nitrogen temperatures and collecting data offers several benefits. Reduction in radiation damage of the crystal on exposure to X-rays, invariably improves the limit of resolution, decreases thermal parameters, allows storage and reuse of crystals, and helps to overcome the scaling problem by enabling the completion of entire data collection using only one crystal (Garman and Schneider, 1997). During data acquisition the crystals are oscillated about an axis perpendicular to the X-ray beam, with a chosen, relatively small angle of oscillation of 1° per frame. Crystal to detector distance is chosen based on
the longest unit cell dimension, mosaic spread etc so that the intensity spots are well resolved and is approximately equal to the longest crystal cell dimension. The exposure time depends on the quality of crystal and oscillation range - larger the oscillation range, longer the exposure time required.

A.I.4. CCP4 program suite

The CCP4 program suite is the program package most widely used by X-ray crystallographers in structure determination and analysis of macromolecules. The CCP4 suite is an integrated set of programs for protein crystallography developed by close collaboration of crystallographers under an initiative by the UK Biotechnology and Biological Sciences Research Council (formerly the ERC). Some software developed elsewhere is also included. The CCP4 suite contains programs for all aspects of protein crystallography, including data processing, data scaling, Patterson search and refinement, isomorphism and molecular replacement, structure refinement, phase improvement and density modification, and presentation of results.

A.I.5. Data Processing

Indexing, processing, scaling and merging of data are carried out by MOSFLM and SCALA of CCP4 (Collaborative Computational Project, Number 4, 1994) suit program. MOSFLM can process diffraction image from a wide range of detectors and produces output as an MTZ files of reflections indicates with their intensities and standard deviation. This MTZ files is then passed in to other programs of the CCP4 program suit for further data reduction. Auto indexing will be carried out by MOSFLM using spots from selected images. SCALA merges the multiple observations of reflections and produce a file that contains averaged intensities of each
reflection. SCALA requires an input MTZ file containing merged intensities, such as that produced by MOSFLM. This program is used to analysis scales and B factor, R merge, completeness, multiplicity, correlation within and between data set can be done. The program TRUNCATE in the CCP4 suit programs (CCP4, 1994) used to convert a file of averaged intensities to a file containing structure factor amplitude (F) and the original intensities using the procedure of French and Wilson.

A.I.6. Mathew’s number

Once the space group and unit cell dimensions of the crystal are known it is possible to estimate the number of molecules in the crystallographic asymmetric unit and the solvent content of the protein crystals with the knowledge of the molecular weight of protein. The following equations are used (Matthews, 1968).

\[
V_m = \frac{\text{Unit cell volume}}{(\text{Mol.Wt.} \times n \times z)}
\]

\[
V_{solv} = 1 - \left(\frac{1.23}{V_m}\right)
\]

Where \(V_m\) is the Mathew’s number, \(n\) is the number of molecules per asymmetric unit and \(z\) is the Avogadro's number; \(V_{solv}\) is the solvent content of protein crystals.

A.I.7. Structure solution

Multi-wavelength Anomalous Dispersion (MAD), Multiple Isomorphous Replacement (MIR) and Molecular Replacement (MR) are the three methods widely used in protein crystallography to solve structures. MAD technique is currently most popular owing to the recent technical advances made in the field of synchrotron radiation, image plate and CCD detectors for data collection and molecular biology techniques that provide selenium derivative for any protein (Ealick, 2000). MIR technique requires more than one heavy atom derivative.
A.I.8. Molecular replacement (MR)

This technique is the simplest of all and can be used when a homologous protein with structural similarity is available in the database. The pioneering studies of Rossmann and Blow in 1962 laid the foundation of Molecular Replacement method (Rossmann, 2001). Owing to the rapid expansion of Protein Data Bank (PDB; Berman et al., 2000) with the increase in the number of models available, MR method is now routinely used in the determination of protein structures.

Molecular replacement is critically dependent on the quality of the search model. Several model-preparation procedures are integrated in the molecular-replacement program MOLREP (Vagin and Teplyakov, 1997). These include model modification on the basis of amino-acid sequence alignment and model correction based on analysis of the solvent-accessibility of the atoms. MOLREP is an automated program for MR (Vagin and Teplyakov, 1997) where, along with the default protocol, there are various search strategies as options. A special feature of MOLREP is that it offers several built-in model preparation functionalities. The integration of model-preparation and Patterson function techniques in one program has several advantages. Apart from convenience, such integration allows specific adjustment of the model-modification parameters for an efficient Patterson search. Moreover, the weighting parameters for the rotation function (RF) and translation function (TF) functions are more reliable if they are derived from the original sequence and atomic coordinates of the homologous protein. Such an integrated approach has proven to be efficient and has recently been implemented in several MR pipelines including BALBES (Long et al., 2008), MrBUMP (Keegan and Winn, 2008) and JSCG (Schwarzenbacher et al., 2008), in which MOLREP itself is used as a
component. Currently, *MOLREP* is being updated to fulfill the requirements raised by *BALBES* development and benefits from its training. The first stage of model preparation in MOLREP includes alignment of the sequence derived from the search model to the target sequence and, provisionally, deletion of residues and atoms of the search model that do not map on to the target sequence. This is a conservative approach to model correction, as no new atoms are added and the coordinates of preserved atoms are not changed. Besides, the sequence identity derived from the alignment of the two sequences is further used for weighting of the X-ray data.

The sequence alignment implemented in MOLREP is a modified version of the dynamic alignment algorithm (Needleman and Wunsch, 1970) which takes into account the known three-dimensional structure of the search model. Thus, buried residues contribute to the total alignment score more than residues at the surface. In addition, it is assumed that gaps and insertions are impossible within sequence segments corresponding to helices and strands. Producing a search model is only one aspect of model preparation. Another aspect is to define the weighting scheme that is most suitable for a given model. During model preparation, *MOLREP* estimates a number of parameters for the search model. Two of them, the radius of gyration of the model and its sequence identity with the target protein, are translated into the parameters of Gaussian low-pass and high-pass filters which define defining structural details of finer scale than the r.m.s.d. between the search and target molecules. The exact value of this r.m.s.d. is the treatment of translational NCS, otherwise known as pseudo-translation, is a special case of model modification because it is applied to the TF but not to the RF. In *MOLREP*, this modification is applied implicitly in reciprocal space and therefore can be considered as either model
modification or as weighting of the X-ray data. Translational NCS is detected and the NCS translation vector is derived using the experimental Patterson function. *MOLREP* assumes that translational NCS is present if there is a non-origin peak in the Patterson function with a height of 1/8 or more of the origin peak height. An additional requirement is that this peak is sufficiently distant from the origin (three-quarters of the diameter of the search model) to ensure that it is not caused, for example, by regular structural patterns in the target molecule. Such an approach is simple, works in most cases and is therefore used by default. However, neither false positives nor false negatives can be excluded.

**A.I.9. Structure refinement**

**A.I.9.1. Refinement and model building**

An initial model built into an experimental map of approximately phased molecular replacement solution, will usually contain many errors. Generating the correct structure from structure factors and initial phases is composed of two parts. The first part is the computational refinement in which the difference in observed and calculated structure factors (R-factor) and differences between observed and optimal stereo chemical, electrostatic and van der Waal’s contacts are minimized. The second part is manual intervention. Computational refinement tends to converge in a local minimum of the total energy landscape. Manual intervention allows for changes to the structures that are outside the radius of convergence. Refinement of a protein structure involves minimizing a function involving observed ($F_o$) and calculated structure factors ($F_c$) and along with applying restraints such as minimizing the difference between the refined parameters and ideal parameters of the stereo chemistry. Refinement techniques generally fall into two categories, depending upon whether the calculations
are performed either in the real space or in the reciprocal space. Refinement methods based on reciprocal space is preferred over real space because the former ones are computationally less expensive. To prevent the model from going into local minimum, interactive graphics is used for checking the fit of the model to the electron density. To monitor the progress of refinement the crystallographic parameter called R factor is used. This is defined as:

\[ R = \frac{\sum F_o - F_c}{\sum F_o} \]

A.I.9.2. Rigid body refinement

Rigid body refinement is the first step in a protein structure refinement procedure that fixes the gross features of a molecule for further refinement. The adjustment of the model consists of refining the three positional parameters and one temperature factor of all the atoms in the structure except hydrogen atoms. This procedure minimizes the R-factor value by refining three rotational and three translational degrees of freedom of the subunits or groups. It is possible to regard the entire molecule as a rigid entity and refine its position and orientation in the unit cell. Each molecule or subunit is treated as a continuous mass distribution located at the center of mass position defined by,

\[ R_j = \frac{1}{M_j} \sum \rho_i \]

Where \( M_j = \sum m_i \)

The \( m_i \) terms are the atomic masses and \( J \) labels the rigid bodies. Positional refinement is done using constraints and restraints. The ratio of the number of parameters to be determined to the number of observations is very high in a protein structure determination. This ratio could be improved in two ways,
one is by reducing the number of parameters by the use of constraints, and
the other by increasing the number of observations by the addition of
restraints. The stereo chemical information such as bond lengths, bond
angles, conformational angles, planarity etc which are obtained from small
molecule structures, are used for applying restraints and constraints.
Restraints are considered when a specific parameter has restricted freedom,
limited to a range of values, whereas, the parameter is constraint, when it
can assume only a specific value.

A.I.9.3. Restrained refinement

During the course of refinement of macromolecules some groups of atoms
may have to be constrained or restrained to improve the ratio of observables
to parameters. As rigid bodies, or, restrain or constrain the bond lengths,
bond angles, non crystallographic symmetry (NCS) and atomic positions to
a desired value by use of appropriate force constants. Restraints are given
when limited freedom can be given for a parameter. When a parameter has
to be held to an exact value, then it is constrained. In practice, a constraint is
a restraint with infinite force constant. In NCS symmetry restraints, the
molecules in the asymmetric unit are superposed by least squares
superposition and the average coordinates \( x_{av} \) of individual atoms are
computed. If \( x \) represents the coordinates of individual atoms, then each
atom can be restrained according to the mathematical term:

\[
E_{NCS} = w(x-x_{av})^2
\]

The corresponding B-factor restraints are given by:

\[
B_{NCS} = (b-b_{av})^2 / \sigma^2_{NCS}
\]
Where \( w \) is a weight function, \( b \) and \( b_{av} \) are the respective individual and average temperature factors of NCS related atoms and \( \sigma_{NCS} \) is the target deviation for B-factor restraints.

**A.I.9.4. Refinement by maximum-likelihood method**

The initial models from MR calculations are refined using the program REFMAC (implemented in CCP4) which makes use of maximum-likelihood equations. REFMAC (Murshudov, Vagin and Dodson 1997) is a macromolecular refinement program which has been integrated into the CCP4 suite. Read (1990) and Bricogne (1991) had suggested a maximum-likelihood target that should be based on various probability distributions. One of the expected advantages of maximum likelihood refinement is a decrease in refinement bias, as the calculated structure-factor amplitudes will not be forced to match the observed amplitudes (Read, 1997). Use of appropriate likelihood targets through the incorporation of the effect of measurement of error and the use of cross-validation data to estimate the \( \sigma \) (sigma) values are the key ingredients in the likelihood refinement. Verification tests have shown that for refinement, maximum likelihood method is more than twice as effective compared to least-squares method, in improving the model (Pannu and Read, 1996). REFMAC program can carry out rigid body restrained or unrestrained refinement using X-ray data (Murshudov, Vagin and Dodson, 1997) or idealization of a macromolecular structure. There are options to use different minimization methods. The program minimizes the coordinate parameters to satisfy a maximum-likelihood or least squares residual. REFMAC also produces an output file with extension MTZ (named after three of its progenitors, McLaughlin,
Terry and Zelinka) containing weighted coefficients for a weighted mF_o-DF_c and 2mF_o-DF_c maps. About 5% of the reflections were kept aside during refinement to calculate R_free for cross validation (Brunger, 1992). NCS averaging is effectively done for an asymmetric unit composed of N similar objects related by non-crystallographic symmetry (NCS). Rossman and Blow (1963) proposed this method, by which the current phases of reflections can be improved by averaging over the electron densities of NCS related objects. NCS averaging requires an accurate estimate of NCS operators and exact information on the position and shape of the objects whose density has to be averaged (Vellieux and Read, 1997).

A.I.10. Graphics, visualization, electron density maps and model building

After phasing the reflections obtained from X-ray diffraction, using one of the methods discussed above, an electron density map is calculated using Fourier transform. The formula for the Fourier summation to calculate an electron density map is:

\[
\rho(xyz) = \frac{1}{V} \sum \sum |F| \cos[2\pi(hx+ky+lz) - \alpha(hk\ell)]
\]

Where x, y, z are the fractional coordinates of each point in the unit cell, F (hkl) is the structure factor, V is the unit cell volume and \(\alpha(hk\ell)\) is the phase angle. The h, k, l, are Miller indices. The difference maps such as 2F_o- F_c and F_o- F_c for electron density are used to identify errors in the model structure and to refine the positional and displacement parameters. Generally the deviations of the model from the reference molecule could be detected in the F_o-F_c difference map. Poorly defined regions of the map are examined with the maps contoured at lower levels. Difference electron density maps are important for locating bound ligands in protein structures (Glusker, Lewis and Rossi, 1994). The graphic software program COOT is
used for displaying and examining the electron density maps, for displaying atoms, for interactive fitting and optimizing the geometry and also for solvating the structures. The displayed map and the molecule could be rotated and viewed from any direction. Each residue starting from N-terminal to the C terminal is examined for their optimum fitting in the electron density maps and corrects geometry. The deviating ones were corrected using Real Space Refinement. A difference Fourier (F_o-F_c) map was calculated to identify the deviations of the protein from its search model. Both the aligned sequences of the respective families and the observed difference densities were used to ascertain the identity of dissimilar residues. After every cycle of visual fitting of the model to the calculated electron density, it is subjected to a few cycles of refinement using REFMAC. After the refinement the changed residues were carefully checked in the new map. The fitted residue was retained if no difference in density (F_o- F_c) was observed and R-factor and R_free improved. This procedure is repeated a number of times till all the observed difference densities in the vicinity of protein atoms were accounted and both R-factor and R_free values converged. The solvent molecules were placed wherever the F_o- F_c density was observed above 3 σ level and when the water molecules made reasonable hydrogen bonds. To add water molecules, difference map peak analysis is carried out and the map above 5 σ was observed and manual water addition is carried out. It is further analyzed by CHECK WATER in the validate panel and the water molecule with close contact less than 2.3 σ and greater than 3.5 σ and map level 1 σ were either removed or arranged in space to get the exact fitting. Coot water was used for adding water molecules to each of the structures through interactive mode. The addition of water molecules is started only when the refinement
of the model reached an R factor below 25%. The water search is usually limited to distance of 5.0 Å from protein, and the distance between any two waters not less than 2.5 Å. Initially water molecules are added at a higher sigma level, later when the structures were refined; this progressively is reduced to lower levels, up to 3.5 σ. The presence of metal ions and anions such as chloride ions are indicated by very high electron density. Metal ions can be added by appropriate option in coot. They are added after the addition of the probable ligand.
Appendix II

PRINCIPLES OF PROTEIN CRYSTALLIZATION

A.II.1. Introduction

As the term X-ray crystallography implies, the sample is analyzed in the crystalline state. Whereas inorganic crystals can be grown to dimensions of several centimeters, it is impossible to grow protein crystals as large as 1 mm in their shortest dimension. Larger crystals are often twinned and useless for data collection. Protein molecules in the crystals stick to each other primarily by hydrogen bonding through intervening water molecules. Protein crystals have 40 to 60% of solvent content, and are much more fragile compared to inorganic or small organic molecule crystals. Proteins retain their function in the crystalline state and conclusions can be drawn about molecular function of the protein by examining them in the crystalline state. Substrates added to suspensions of crystalline enzymes are converted to products, showing that the catalytic and substrate binding sites are intact.

The process of crystallization is the limiting step in protein crystallography. The principles of crystallization of macromolecules are analogous to those of small molecules. Protein molecules themselves are distinctive, composed of approximately 50% solvent though this may vary from 30 – 78 % (Matthews, 1985). Protein crystals are labile, fragile and sensitive to external environments owing to their high solvent content and weak binding energies between protein molecules in the crystal. The only optimal conditions suitable for their growth are those that cause little or no perturbation of their molecular properties. As a result, crystals must be grown from a medium where temperature is constant and within a broad range of pH 3 to 10 because complete hydration is an essential factor for maintaining the crystal structure integrity. There are many differences
between small molecule crystals and protein crystals. In general, small molecule crystals are grown to large dimensions. They are physically hard, brittle, easy to handle, have strong optical properties and can diffract X-rays strongly. This is because the small molecule crystals exhibit firm lattice interactions and highly ordered lattice arrangements within the crystal. In comparison, protein crystals are generally smaller in size (1 to 1000 µm), soft and crushed without difficulty and with weak crystal forces. These kinds of crystals, which will redissolve if rehydration occurs, have weak optical properties and diffract X-rays weakly. They are also temperature sensitive, because protein solubility alters as a function of temperature. In theory, as the temperature rises, the solubility of proteins increases, thus, no crystals are formed or crystals get redissolved. When crystals are exposed to long-standing X-ray radiation, it can cause extensive damage of crystals. This is due to the weakening of lattice forces within the crystal structure (McPherson, 1982) and the X-rays can produce sufficient free radicals to cause specific chemical changes on the protein molecules such as the breaking of the disulfide bonds (Ravelli and McSweeney, 2000).

The crystallization of proteins from solution is a reversible equilibrium phenomenon. It contains three stages: nucleation, growth and cessation of growth. The formation of crystals is due to the decreasing free energy of the system while the formation of many new chemical bonds simultaneously outweighs the decreasing entropy of the system in order to grow a highly organized internal structure. In other words, the free energy of the system is reduced to its energy minimum and a thermodynamic driving force exists that provides for the ordering process of crystals (McPherson, 1982). The basic strategy of producing protein crystals is to generate a certain degree of super saturation in the solution. At the equilibrium point, the amount of
protein molecules entering the solution is the same as the amount of protein molecules leaving the solution. This is referred to as the solubility limit of a protein. When the solubility of a protein is below this limit, the solution is under saturated. If the solubility is equal to the limit, the solution is saturated. Crystals can grow only when the solubility exceeds the limit. Every protein has a unique solubility. Decreasing the solubility of the protein is the most effective way to create supersaturation for crystal growth. Only in a non-equilibrium supersaturated solution, can a crystal grow. Supersaturation can be achieved by different approaches including altering the buffer pH, temperature, protein concentration, dielectric constant of the medium and precipitant concentration in order to change the protein solubility to reach the condition that lies just above the supersaturation region (McPherson, 1998).
A classical explanation of crystal nuclei formation and growth can be visualized by the two-dimensional solubility phase diagram shown in the figure below. The solubility curve divides the concentration space into under saturation and super saturation regions. In the under saturation zone, under the solubility curve, the protein will never crystallize. Above the solubility curve, this region can be subdivided into three zones according to level of saturation and the kinetics required for reaching equilibrium. In the precipitation zone, excess protein does not remain in solution and exists as an amorphous precipitate the formation of precipitate implies that no crystals will form. Before a crystal can grow in solution, nucleation has to occur. Nucleation is the beginning of crystal formation. In this process, the nucleus of sufficient size must be formed to initiate aggregation in an ordered manner. In the nucleation zone or labile zone, there is a high probability that critical nuclei will form spontaneously in solution because this corresponds to an increased energy state of the system.
In fact, the energy or probability barrier to the formation of the first nucleus allows the creation of a supersaturated solution. If the degree of aggregation is too high, the solution will be oversaturated and a precipitate will be formed. If the degree of aggregation is adequate, stable nuclei can continue to grow to larger size without forming precipitate, and then the crystal can be formed and grown. The detestable zone is ideal for the growth of crystals without nucleation of new crystals. In this zone nuclei will not form, but if nuclei are present or seed crystals are introduced then crystals may grow. When a crystal grows to a certain size, it will stop growing spontaneously. The crystallization of proteins is influenced by a number of factors, and each protein is unique. It is not possible to envisage the conditions that can cause the crystallization of a protein. The various parameters that affect crystallization are not independent of each other and their interrelation may be complicated and difficult to distinguish (McPherson, 1998). Finding a rational guideline to crystallize macromolecules successfully is not an easy task. The only way to do this is to identify the important components and refine each of them individually. In general, precipitant type and concentration, buffer type and pH, temperature, and sample concentration are the most important factors for protein crystallization. They are considered first when performing crystallization experiments. Each parameter is manipulated independently to determine its effect on crystallization.

A.II.2. Crystallization Methods

There are at least seven practical methods used for macromolecule crystallization including vapor diffusion, bulk crystallization, batch, free-interface diffusion, dialysis, temperature-induced and seeding. Among these,
vapor diffusion and micro batch methods are the most popular means being utilized by crystallographers to obtain macromolecular crystals. The method of vapor diffusion is undoubtedly regarded as the most widely employed approach for crystallization. Nucleation occurs when the sample concentration increases as the droplet volume decreases by hydration-driven mechanisms. This is induced by the equilibration of water vapor between the sample droplet and the reservoir solution. The vapor diffusion technique is an ideal methodology for screening a broad spectrum of crystallization conditions. It can be used to optimize the size of crystals suitable for X-ray diffraction analysis. Vapor diffusion methods include hanging drop, sitting drop, sandwich, and capillary methods. The most common protocols are the hanging drop and sitting drop methods (McPherson, 1998). The hanging drop vapor diffusion method is an efficient means of screening crystallization parameters. The advantage of this method is that it requires only a small volume of droplet, which can be as low as 2 µL per experiment, so a minimum amount of sample is consumed for screening and optimization of the crystallization conditions (McPherson, 1998). The reason for the popularity of the hanging drop method is the ease of performing the experiment, only a 24 well cell culture plate, silicone grease and cover slips siliconized with dichloro dimethyl silane are needed. The principle of this approach is straightforward, a drop composed of a mixture of macromolecule sample and precipitating solution is placed in vapor equilibration with a reservoir solution of precipitant and buffer. To start the trial, the precipitating solution composed of precipitant, buffer, additive etc, is dispensed into the reservoir. Then equal volumes of the sample and reservoir solution are mixed onto the surface of siliconized glass cover slide. The drop has a lower concentration of precipitant than the reservoir solution,
so water or volatile chemicals will escape from the drop into the reservoir solution to achieve system equilibrium inside the reservoir. Eventually this causes the sample inside the drop to become more concentrated until the precipitating concentration in the drop is almost equivalent to the reservoir concentration. The major benefits of using the hanging drop method are relative ease of mounting the crystal for X-ray diffraction experiments by inverting the cover slide with a pair of forceps. This method can be used to place multiple drops in each reservoir as well, thus saving time and material. Frequently, very small crystals are obtained instead of large ones. Small crystals of good quality are useful as seeds to grow large crystals. Each hanging drop is seeded with a small number of small crystals. Crystals may grow from seeds nearly 10 times faster than they grow anew, so most of the protein goes into a few crystals, increasing their size.

Co-crystallization is a process in which crystals of protein-ligand complexes are grown by crystallizing protein and ligand together. This is the only method for producing crystals of proteins in complexes with large ligands, such as nucleic acids or other proteins. Another means of obtaining complex crystals is to soak protein crystals in the mother liquor containing the ligand. This is practical in the case of small ligands alone. Proteins retain their activity in the crystalline state, and ligands can diffuse to the active sites and binding sites through channels of water in the crystal. Soaking is preferred usually over cocrystallisation for comparison of structure of native protein with that of protein ligand complex. Soaking of preformed protein crystals with ligands is more likely to produce crystals of same form and unit cell dimensions as those of the native protein (Rhodes, 2000).

Many variables influence the formation of protein crystals. These include obvious ones like purity of protein, concentration, precipitant concentration,
pH, temperature as well as more subtle ones like cleanliness, vibration, source and age of the protein and the presence of ligands (David, 2000).
Appendix III

MOLECULAR DOCKING

A.III.1. Introduction

In the field of molecular modeling, docking is a method which predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex. Knowledge of the preferred orientation in turn may be used to predict the strength of association or binding affinity between two molecules using for example scoring functions (Lengauer and Rarey, 1996).

The associations between biologically relevant molecules such as proteins, nucleic acids, carbohydrates, and lipids play a central role in signal transduction. Furthermore, the relative orientation of the two interacting partners may affect the type of signal produced (e.g., agonism vs antagonism). Therefore docking is useful for predicting both the strength and type of signal produced.

Docking is frequently used to predict the binding orientation of small molecule drug candidates to their protein targets in order to in turn predict the affinity and activity of the small molecule. Hence docking plays an important role in the rational design of drugs (Kitchen et al., 2004). Given the biological and pharmaceutical significance of molecular docking, considerable efforts have been directed towards improving the methods used to predict docking.

Molecular docking can be thought of as a problem of “lock-and-key”, where one is interested in finding the correct relative orientation of the “key” which will open up the “lock” (where on the surface of the lock is the key hole, which direction to turn the key after it is inserted, etc.). Here, the protein can be thought of as the “lock” and the ligand can be thought of as a
“key”. Molecular docking may be defined as an optimization problem, which would describe the “best-fit” orientation of a ligand that binds to a particular protein of interest. However since both the ligand and the protein are flexible, a “hand-in-glove” analogy is more appropriate than “lock-and key (Jorgensen, 1991). During the course of the process, the ligand and the protein adjust their conformation to achieve an overall “best-fit” and this kind of conformational adjustments resulting in the overall binding is referred to as “induced-fit (Wei et al., 2004). The focus of molecular docking is to computationally stimulate the molecular recognition process.

The aim of molecular docking is to achieve an optimized conformation for both the protein and ligand and relative orientation between protein and ligand such that the free energy of the overall system is minimized.

A.III.2. Docking approaches

Two approaches are particularly popular within the molecular docking community. One approach uses a matching technique that describes the protein and the ligand as complementary surfaces. The second approach simulates the actual docking process in which the ligand-protein pairwise interaction energies are calculated. Both approaches have significant advantages as well as some limitations. These are outlined below.

A.III.3. Shape complementarity

Geometric matching/ shape complementarity methods describe the protein and ligand as a set of features that make them dockable. These features may include molecular surface/ complementary surface descriptors. In this case, the receptor’s molecular surface is described in terms of its solvent-accessible surface area and the ligand molecular surface is described in terms of its matching surface description (Shoichet, Kuntz, Bodian, 2004). The complementarity between the two surfaces amounts to the shape
matching description that may help finding the complementary pose of
docking the target and the ligand molecules. Another approach is to describe
the hydrophobic features of the protein using turns in the main-chain atoms.
Yet another approach is to use a Fourier shape descriptor technique. Whereas
the shape complementarity based approaches are typically fast and
robust, they cannot usually model the movements or dynamic changes in the
ligand/ protein conformations accurately, although recent developments
allow these methods to investigate ligand flexibility. Shape complementarity
methods can quickly scan through several thousand ligands in a matter of
seconds and actually figure out whether they can bind at the protein’s active
site, and are usually scalable to even protein-protein interactions. They are
also much more amenable to pharmacophore based approaches, since they
use geometric descriptions of the ligands to find optimal binding (Kahraman
et al., 2004).

A. III. 4. Simulation

The simulation of the docking process as such is a much more complicated
process. In this approach, the protein and the ligand are separated by some
physical distance, and the ligand finds its position into the protein’s active
site after a certain number of “moves” in its conformational space. The
moves incorporate rigid body transformations such as translations and
rotations, as well as internal changes to the ligand structure including torsion
angle rotations. Each of these moves in the conformation space of the ligand
induces a total energetic cost of the system, and hence after every move the
total energy of the system is calculated. The obvious advantage of the
method is that it is more amenable to incorporate ligand flexibility into its
modeling whereas shape complementarity techniques have to use some
ingenious methods to incorporate flexibility in ligands. Another advantage is
that the process is physically closer to what happens in reality, when the protein and ligand approach each other after molecular recognition. A clear disadvantage of this technique is that it takes longer time to evaluate the optimal pose of binding since they have to explore a rather large energy landscape. However grid-based techniques as well as fast optimization methods have significantly ameliorated these problems.

A.III.5. Mechanics of docking

To perform a docking screen, the first requirement is a structure of the protein of interest. Usually the structure has been determined using a biophysical technique such as X-ray crystallography, or less often, NMR spectroscopy. This protein structure and a database of potential ligands serve as inputs to a docking program. The success of a docking program depends on two components: the search algorithm and the scoring function.

A.III.6. Search algorithm

The search space consists of all possible orientations and conformations of the protein paired with the ligand. With present computing resources, it is impossible to exhaustively explore the search space—this would involve enumerating all possible distortions of each molecule (molecules are dynamic and exist in an ensemble of conformational states) and all possible rotational and translational orientations of the ligand relative to the protein at a given level of granularity. Most docking programs in use account for a flexible ligand, and several are attempting to model a flexible protein receptor. Each "snapshot" of the pair is referred to as a pose. There are many strategies for sampling the search space.

- Use a coarse-grained molecular dynamics simulation to propose energetically reasonable poses
• Use a "linear combination" of multiple structures determined for the same protein to emulate receptor flexibility
• Use a genetic algorithm to "evolve" new poses that are successively more and more likely to represent favorable binding interactions.

A.III.7. Scoring function

The scoring function takes a pose as input and returns a number indicating the likelihood that the pose represents a favorable binding interaction. Most scoring functions are physics-based molecular mechanics force fields that estimate the energy of the pose; a low (negative) energy indicates a stable system and thus a likely binding interaction. An alternative approach is to derive a statistical potential for interactions from a large database of protein-ligand complexes, such as the Protein Data Bank, and evaluate the fit of the pose according to this inferred potential.

There are a large number of structures from X-ray crystallography for complexes between proteins and high affinity ligands, but comparatively fewer for low affinity ligands as the later complexes tend to be less stable and therefore more difficult to crystallize. Scoring functions trained with this data can dock high affinity ligands correctly, but they will also give plausible docked conformations for ligands that do not bind. This gives a large number of false positive hits, i.e., ligands predicted to bind to the protein that actually do not bind when placed together in a test tube.

One way to reduce the number of false positives is to recalculate the energy of the top scoring poses using (potentially) more accurate but computationally more intensive techniques such as Generalized Born or Poisson-Boltzmann methods (Feig et al., 2004).
A.III.8. Applications

A binding interaction between a small molecule ligand and an enzyme protein may result in activation or inhibition of the enzyme. If the protein is a receptor, ligand binding may result in agonism or antagonism. Docking is most commonly used in the field of drug design — most drugs are small organic molecules, and docking may be applied to:

• hit identification – docking combined with a scoring function can be used to quickly screen large databases of potential drugs in silico to identify molecules that are likely to bind to protein target of interest (virtual screening).

• lead optimization – docking can be used to predict in where and in which relative orientation a ligand binds to a protein (also referred to as the binding mode or pose). This information may in turn be used to design more potent and selective analogs.

• Bioremediation – Protein ligand docking can also be used to predict pollutants that can be degraded by enzymes.

In the present study, molecular docking was conducted to assess the suitability of the bioconverted derivatives of berberine as PLA2 inhibitors and compare the binding strength of berberine with its derivatives. Binding affinities were monitored by using the Glide-score values. The binding affinities and the mode of binding of both the native and biotransformed forms of berberine were compared. Also, docking was conducted with various types of PLA2s with berberine.
APPENDIX-IV

List of PDB entries

PDB entries used in this dissertation.

1. 1FV0: Structural basis of phospholipase A$_2$ inhibition for the synthesis of prostaglandins by the plant alkaloid aristolochic acid from a 1.7 A crystal structure.
   Chandra, V., Jasti, J., Kaur, P., Srinivasan, A., Betzel, C., Singh, T.P.

2. 1O2E: Crystal structures of the free and anisic acid bound triple mutant of phospholipase A$_2$.
   Sekar, K., Vaijayanthi Mala, S., Yogavel, M., Velmurugan, D., Poi, M.J., Vishwanath, B.S., Gowda, T.V., Jeyaprakash, A.A., Tsai, M.D.

3. 2QU9: Crystal structure of the complex of group II phospholipase A$_2$ with Eugenol
   Journal: To be Published 2007.

4. 1Q7A: Phospholipase A$_2$ as a target protein for nonsteroidal anti-inflammatory drugs (NSAIDS): crystal structure of the complex formed between phospholipase A$_2$ and oxyphenbutazone at 1.6 A resolution.
   Singh, N., Jabeen, T., Somvanshi, R.K., Sharma, S., Dey, S., Singh, T.P.
5. **2B17**: Specific binding of non-steroidal anti-inflammatory drugs (NSAIDs) to phospholipase A$_2$: structure of the complex formed between phospholipase A$_2$ and diclofenac at 2.7 Å resolution.


6. **2OTF**: Crystal structure of the complex formed between phospholipase A$_2$ and atenolol at 1.95 Å resolution

Kumar, S., Singh, N., Sharma, S., Bhushan, A., Kaur, P., Singh, T.P.


7. **2QUE**: Saturation of substrate-binding site using two natural ligands: Crystal structure of a ternary complex of phospholipase A$_2$ with anisic acid and ajmaline at 2.25 Å resolution

Kumar, S., Singh, N., Sharma, S., Kaur, P., Singh, T.P.

**Journal**: To be Published 2007

8. **2OTH**: Crystal structure of a ternary complex of phospholipase A$_2$ with indomethacin and nimesulide at 2.9 Å resolution

Kumar, S., Singh, N., Sharma, S., Kaur, P., Singh, T.P.

**Journal**: To be Published 2007

9. **1TGM**: Crystal structure of a complex formed between group II phospholipase A$_2$ and aspirin at 1.86 Å resolution

Singh, N., Jabeen, T., Sharma, S., Bhushan, A., Singh, T.P.

**Journal**: To be Published 2004.

11. 3KK6: Coxibs interfere with the action of aspirin by binding tightly to one monomer of cyclooxygenase-1.
13. 4P2P: An independent crystallographic refinement of porcine phospholipase A$_2$ at 2.4 A resolution.
Finzel, B.C., Ohlendorf, D.H., Weber, P.C., Salemme, F.R.
15. 1FB2: Regulation of catalytic function by molecular association: structure of phospholipase A$_2$ from Daboia russelli pulchella (DPLA$_2$) at 1.9 A resolution.
Chandra, V., Kaur, P., Jasti, J., Betzel, C., Singh, T.P.
16. **1LWB**: Atomic resolution structure of prokaryotic phospholipase A2: analysis of internal motion and implication for a catalytic mechanism.

   Matoba, Y., Sugiyama, M.


17. **1KVO**: High-resolution X-ray crystallography reveals precise binding interactions between human nonpancreatic secreted phospholipase A2 and a highly potent inhibitor (FPL67047XX).


Appendix V

STRUCTURE ALBUM

- Arachidonic Acid
- Ajmaline
- Anisic Acid
- Aspirin
- Aristolochic Acid
- Atanolol
- Berbemuline
- Berberine
- Columbamine