CHAPTER 2

Macromolecular Crystallographic Method and Protein Structure Determination

2.1 Principles of X-ray crystallography 48
2.2 Tools for X-ray diffraction 49
  2.2.1 X-ray sources 49
  2.2.2 X-ray detectors 53
  2.2.3 Other methods to study 3-D structure of macromolecules 53
2.3 Single Crystal Diffraction Data Collection 55
  2.3.1 Oscillation Method 55
2.4 Data Processing 56
2.5 Phase Solution 58
  2.5.1 Molecular Replacement (MR) method 59
    2.5.1.1 The Rotation Function 59
    2.5.1.2 The Translation Function 60
    2.5.1.3 Maximum Likelihood based MR method 61
  2.5.2 Multiple Isomorphous Replacement (MIR) 63
  2.5.3 Multiple Wavelength Anomalous Diffraction (MAD) 64
  2.5.4 Direct Methods 64
  2.5.5 Other phasing method 65
2.6 Density modification 66
2.7 Electron density map interpretation 66
2.8 Crystallographic refinement 67
  2.8.1 Target functions 67
  2.8.2 Maximum Likelihood refinement targets 68
  2.8.3 Minimization method - Simulated Annealing 69
  2.8.4 Multi-start refinement 70
  2.8.5 TLS refinement 70
  2.8.6 R-factor and cross-validation 70
2.9 Validation of the model 71
2.1 Principles of X-ray crystallography

Single crystal diffraction is a very powerful method of determining accurately three dimensional structures of molecules of any size. A single crystal is a periodic array of identical unit cells repeating translationally on a three-dimensional lattice defined by the basis vectors of the unit cell. If there are atoms in the unit cell, the crystal can be regarded as made up of N interpenetrating lattices of identical basis vectors, but with origins displaced by positional vectors of each of the N-atoms.

Incident monochromatic X-rays are resonantly scattered by the oscillating free electrons of the atoms in the crystal [199]. The long range periodicity of the lattice in three dimensions leads to constructive interference resulting in non-zero intensities of the diffracted secondary X-rays only along specific directions. This is mathematically expressed as the three Laue diffraction conditions [200], which have to be satisfied simultaneously,

\[ \mathbf{S} \cdot \mathbf{a} = h \]
\[ \mathbf{S} \cdot \mathbf{b} = k \]
\[ \mathbf{S} \cdot \mathbf{c} = l \]

These Laue conditions lead to Bragg’s law in reflection geometry [201],

\[ 2d(hkl)\sin\theta = \lambda , \]

where \( \mathbf{a} \), \( \mathbf{b} \) and \( \mathbf{c} \) are the crystal translation vectors, \( \mathbf{S} \) is the scattering vector with magnitude \( 2\sin\theta/\lambda \), \( h \), \( k \), \( l \) are the Miller indices of the reflecting lattice planes, \( d \) is the distance between the lattice planes \( (hkl) \), \( \theta \) is the angle of reflection and \( \lambda \) is the incident X-ray wavelength. Considering the crystal to be made up of interpenetrating lattices, formed by each atom of the unit cell occupying the origin of the respective lattice, the structure factor, \( \tilde{F}(hkl) \), can be written as,

\[ \tilde{F}(hkl) = \sum_{j=1}^{N} f_j e^{2\pi i (h\mathbf{r}_j + k\mathbf{r}_j + l\mathbf{r}_j)} \]

where \( f_j \) is the atomic form factor of the \( j \)th atom positioned at \( \mathbf{r}_j \). Treating electron density as a continuous function \( \rho(xyz) \) with peaks at atomic positions \( \mathbf{r}_j \), the the structure factor is written as

\[ \tilde{F}(hkl) = V \int_{cell} \rho(xyz) e^{2\pi i (hx + ky + lz)} \, dx dy dz \]

where \( V \) is the volume of the unit cell and \( \rho \) is the electron density distribution. Inverse Fourier transformation of equation 4 gives the electron density distribution in the unit cell,
\[
\rho(xyz) = \frac{1}{V} \sum_h \sum_k \sum_l \tilde{F}(hkl) e^{-2\pi i (hx+kx+lx)}
\] (5)

The structure factor of equation (4) is a complex quantity, and is therefore characterized by an amplitude, \( |\tilde{F}(hkl)| \) and a phase \( \varphi(hkl) \).

\[
\tilde{F}(hkl) = |\tilde{F}(hkl)| e^{i\varphi(hkl)}
\] (6)

The amplitudes \( |\tilde{F}(hkl)| \) can be derived from the experimentally measured intensities of the diffracted beam \( I(hkl) \), but the phase angles \( \varphi(hkl) \) cannot be obtained directly from the diffraction pattern. Without phase angle, equation 5 cannot be solved, and this is known as 'phase problem' in crystallography [202].

2.2 Tools for X-ray diffraction

2.2.1 X-ray sources

X-radiation or X-rays is a form of electromagnetic radiation having wavelength in the range of 0.01 to 10 nanometers and energies in the range 120 eV to 120 keV. There are two types of X-ray sources used commonly by the crystallographer today, the conventional laboratory X-ray generators and the synchrotrons. A conventional laboratory source consists of an X-ray tube in which the electrons released from a cathode (tungsten filament) are accelerated in high vacuum by high applied voltage. The high velocity electrons collide with a metal target, the anode, creating the X-rays characteristic of the anode [203]. In synchrotrons (Figure 12a) charged particles (electrons or positrons) are accelerated in a storage ring at nearly the speed of light producing electromagnetic radiation by bremsstrahlung. Since freely traveling electrons (positrons) are not quantized, the emitted radiation ranges over a wide wavelength region, depending on the energy of the charged particles and on the strength of the magnetic field inside the storage ring [204]. The tangentially emitted radiations are used as X-ray sources at synchrotron beamlines (Figure 12b). A useful characteristic is the median of the distribution of power over the spectral region, called the critical photon energy \( E_c \), which divides the emitted power spectrum in two equal parts,

\[
E_c = 0.665E^2B
\] (7)

and

\[
\lambda_c = \frac{18.64}{E^2B}
\] (8)
where $E_c$ is in $keV$ and $E$ is the circulating power in $GeV$ given by,

$$E = \frac{Particle\ Energy \times Current}{Revolution\ frequency},$$

(9)

**Figure 12:** a) The European Radiation Synchrotron Facility (ESRF) at Grenoble, France. b) A view of the BM30A synchrotron beamline where diffraction data was collected.

$B$ is the magnetic field strength in Tesla and $\lambda_c$ is in Å. The main photon flux is close to $E_c$, but above $E_c$ it drops exponentially as a function of the photon energy. The magnetic field causes the charged particles to change their direction. There are four types of magnetic devices in storage rings.

a) **Bending magnets** guide the electrons in the storage ring to stay in a circular closed orbit. At each bend the beam is guided into a straight section.

b) **Wavelength shifter** increases the magnetic field strength, $B$ locally resulting in a decrease of $\lambda_c$ (Equation 8) and produces higher intensity X-rays at shorter wavelength.

c) **Multipole wiggler** is a series of wavelength shifters where the radiation from consecutive magnets being independent, adds up incoherently in the general direction of propagation of the electron beam. The total flux is $N$ times the flux generated by a single period (i.e. bending magnet), where $N$ is the number of periods. It is easily tunable to desired wavelength.

d) **Undulators** are similar to multipole wigglers but with moderate magnetic fields and a large number of poles. The effect of this difference is that in the undulator strong interference occurs between the radiation from the consecutive magnets, which results in a spectral profile with a peak at a specific wavelength and a few harmonics. Therefore this radiation is not polychromatic, but the advantage for
some experiments is that a monochromator is not required. Its emitted intensity can be \(N^2\) times that of a single period, with \(N\) the number of poles.

**Figure 13**: a) Comparison of the brilliance between the 3rd generation synchrotron sources and XFELs. b) A prototype of Compact Light Source [214].

Synchrotron sources have evolved through three generations and currently there are about 63 synchrotrons in the world. The advantages of synchrotron radiation for X-ray diffraction are: 1) its high intensity that is about five to six orders of magnitude higher than that from the laboratory-based rotating anode X-ray generators, 2) low divergence of the beam, resulting in sharper diffraction spots, 3) very good energy resolutions (\(\sim 10^{-4}\)) of the X-ray photon energies or wavelengths produced and 4) tunability of X-rays to obtain desired wavelengths to solve the “phase problem” by multiple or single wavelength anomalous dispersion methods [205]. Synchrotron radiation is obtained in picosecond flashes produced by circulating bunches of charged particles. This property is being utilized for time resolved measurements in microsecond range towards understanding enzyme catalysis [206]. Together with the increasing availability of high intensity X-rays from synchrotrons and new crystal flash freezing techniques, data collection on very small crystals and on molecular assemblies has become feasible [207]. Synchrotron beams can be circularly polarized and can be used for circular dichroism studies to study protein-protein complexes [208].

The most promising directions for fourth generation sources in the wavelength range from the VUV to hard X-rays, are storage rings with even lower emittance than third generation rings, and short wavelength free-electron lasers (FELs) which offer femtosecond pulses with full transverse coherence.
The brilliance of these beams can be a billion times more than that of a third generation undulator source (Figure 13a). However, technological hurdles are still there in actual feasibility of operating the FELs in 1.0 Å wavelength regions. There are about nine FEL facilities around the world today. In a recent study, using nanocrystals of photosystem I, one of the largest membrane protein complexes, more than 3,000,000 diffraction patterns were collected, and a three-dimensional data set was assembled from individual photosystem I nanocrystals [211-213]. The problem of radiation damage in crystallography was taken care of by using pulses briefer than the timescale of most damage processes. This offers a new approach to structure determination of macromolecules that do not yield crystals of sufficient size for studies using conventional radiation sources or are particularly sensitive to radiation damage. Future beamlines worldwide are aiming to achieve up to a $10^5$-fold increase in pulse irradiance by tighter focusing, allowing data collection with low-fluence, 10 fs pulses or pulses of even shorter duration. This provides a route to further reducing radiation damage and may allow measurements on even smaller nanocrystals, down to a single unit cell or a single molecule [211]. In future it has opened up the possibility of elucidation of structure from single molecules without the need of crystals!

Another new development is to produce a new generation of compact X-ray sources - miniature synchrotron light sources (Figure 13b). At Lyncean Technologies Inc., USA, the Compact Light Source, delivers a monochromatic beam of tunable hard X-rays, from a few keV up to 35 keV, comparable in quality to some of the most productive synchrotron beamlines in use today [214]. A laser beam colliding with an opposing electron beam has the same effect as an electron beam passing through an undulator magnet. The electric and magnetic fields of the laser beam cause the electron to wiggle and induces a radiation spectrum similar to that from a long undulator magnet. This radiation is typically referred to as Compton Scattering or Inverse Compton Scattering. If a laser beam with a wavelength of one micron is used, the electron beam energy necessary for 1 Å radiation is only about 25 MeV. The X-ray spectrum of the laser pulse/electron bunch collision is equivalent to that of a 20000 period undulator magnet. The X-rays can be focused using conventional X-ray optics down to a size of about 60 microns. Fine-tuning of the X-ray energy (for scans near absorption edges) is achieved with a monochromator adjustment just as with a synchrotron beamline. The instrument "Beta CLS" has been installed and commissioned at Accelerated Technology Center for Gene to 3D
Structure (ATCG3D). In June 2009, a 2 Å dataset on *Mycobacterium tuberculosis* glycine cleavage system protein crystal was collected (PDBID: 3IFT) [215].

### 2.2.2 X-ray detectors

X-ray detection for the protein crystallographic applications are carried out by integrating or single photon counting type detection system. Integrating detectors like image plates and CCDs [216, 217] are mostly used. Due to their large active surface, good spatial resolution and high dynamic range, IP detectors are popular in protein crystallography laboratories across the world [218]. However, the main drawback of IPs is its long readout time (1 min - 8 min), which becomes a problem at synchrotron sources, where exposure times are of the order of seconds. The main advantage of CCDs over IPs is their extremely small readout times (∼ a few seconds) well suited at the synchrotron sources [219, 220]. CCDs achieve sensitivity or a quantum efficiency of about 80% as compared to only a few percent in case of IPs. The typical dynamic range is about $10^5$, which is similar to the IPs. The main disadvantage of CCDs is that they are physically much smaller than IPs and consequently an array of such devices are required for practical applications. However, fabricating a large array of such devices is technologically challenging as the CCDs need to be cooled and joined properly.

A completely new type of detector for protein crystallography applications is a solid-state pixel array detector that offers a huge performance advantage over IPs and CCDs. The detector consists of an array of reverse-biased Si diode array bump-bonded to an Application Specific Integrated Circuit (ASIC) [221-223]. The hybrid technique of bump-bonding gives many advantages. The individual channels have low leakage current and low input capacitance allowing low readout noise. It can be operated in true single photon counting mode without dead-time and provide imaging at high speed. It offers frameless readout for high-speed and time-resolved crystallography, especially when the direct conversion type is used. Such detectors, called PILATUS, are being used at synchrotrons.

### 2.2.3 Other methods to study 3-D structure of macromolecules

The necessity to determine the three dimensional structures of proteins has led to the development of various sophisticated tools based on sound scientific principles. Following four approaches are currently being used to study the structures at a resolution range where individual macromolecules and their internal structure could be visualized.
1. **Neutron Crystallography:** Hydrogen atoms are ubiquitous in a functioning protein and play a vital role in molecular recognition via hydrogen bonds, and hence their location is very important to understanding biomolecular function [224-226]. Exchangeable hydrogens can be precisely located by the neutron diffraction technique, which is similar to X-ray diffraction in principle [227]. Neutron sources are specialized and costly and are of two types: nuclear reactor based (fission neutrons) and accelerator based (spallation neutrons). The neutron data can be coupled to X-ray data in order to refine the atomic positions more accurately and obtain the charge separation in polarized bonds [228].

2. **Cryo-electron microscopy:** In this technique, the biological sample is flash frozen to liquid nitrogen temperatures and investigated in high vacuum by high-energy electrons (typically 200-300 keV) using the principles of transmission electron microscopy (TEM) [229]. In this direct imaging technique 3-D structure is reconstructed from several 2-D projections of the molecule. Cryo-electron microscopy becomes most useful for large assemblies showing them in their native environment in comparison to X-ray crystallography [230]. However there are problems with harsh environment due to high vacuum and sample damage, which means that very low electron dose must be used to avoid destroying the sample, so that the images have extremely poor signal-to-noise ratio and must be averaged out.

3. **Nuclear Magnetic Resonance:** NMR detects chemical shifts of atomic nuclei with non-zero spin (e.g. $^1$H, $^{13}$C and $^{15}$N) [231]. The shifts depend on the electronic environments of the nuclei, namely, the identities and distances of nearby atoms. From these constraints, three-dimensional structure of proteins is derived as a set of 10-50 probable models. About 16% of the structures deposited in the PDB have been determined by NMR. NMR allows the study of the proteins in solution. The protein must however be soluble to a high concentration (0.2-1.0 mM) and be stable for days without aggregation under the experimental conditions. These limitations make the structure elucidation of typical membrane proteins and filamentous proteins unlikely. But intrinsically unstructured proteins, not amenable to crystallization, can be studied by NMR [232]. In principle the size limitation can be overcome by solid state NMR [233].

4. **Small angle scattering:** The scattering peak near the direct beam, below 10°, of a macromolecular solution contains the size and shape information about the biomolecule in solution [234]. The probe can be X-rays or thermal neutrons. This technique is useful to identify the
homogeneity, multimeric states, complexes and foldability of the sample [235, 236]. Low resolution structures can be determined, especially for proteins difficult to crystallize. Contrast matching and upcoming data analysis and refinement programs shall enable low resolution structure determination of multimeric proteins, protein-DNA/RNA and protein-protein complexes in solution [237].

5. **Powder diffraction:** Powder diffraction is an emerging tool to elucidate the structures of proteins from polycrystalline powder samples [238, 239]. This technique is useful when the protein crystals show poor diffraction, are radiation sensitive and form weak complexes [240].

### 2.3 Single Crystal Diffraction Data Collection

Different strategies are used to record within reasonable time and accuracy intensities of all Bragg reflections produced by the crystal. The most common method is to change the orientation of the crystal relative to the incoming monochromatic X-ray beam. This causes a corresponding rotation of the reciprocal lattice and in this manner all reflections can be passed through the Ewald sphere. It is however also possible to change the Ewald sphere radius by changing the wavelength and this is the principle of Laue diffraction. The methods of data collection are given below.

- Oscillation method (single axis rotating crystal, stationery 2D detector, overlap problem) [241]
- Weissenberg method (single axis rotating crystal, moving 2D detector, gives distorted picture of reciprocal lattice) [242]
- Precession method (single axis rotating crystal, moving 2D detector synchronized with the crystal movement, gives undistorted picture of reciprocal lattice) [243]
- Multi-circle diffraction method (multiple axis rotating crystal, moving 1D detector, one reflection at a time, more accurate but relatively slow) [203]
- Laue method (all wavelengths and many reflections at a time, fast, overlap problem) [244]

#### 2.3.1 Oscillation Method

The oscillation method, which has been used for data collection in the present thesis work, will be briefly discussed. With the advent of 2D detectors and their fast response, currently, oscillation method is the most popular method. The oscillation method is a straightforward technique where the crystal is rotated around an axis (the oscillation axis) that is perpendicular to the incident X-ray beam (Figure 14). To avoid overlapping of reflections on the stationary 2D detector, the crystal can be
rotated over only a small angle (0.25°-2°) per diffraction image. Crystal is oscillated up and down this angular range several times (called passes) during each exposure, ensuring changes in beam intensity are averaged out. Depending upon the spot overlaps the oscillation width needs to be reduced. The total rotation should be 180° in order to collect all diffracted rays. However, there is redundancy in the reciprocal lattice due to point group symmetry, and hence rotation range is determined by 180°/n (where ‘n’ is the symmetry number).

![Diagram of X-ray diffraction experiment]

**Figure 14:** A schematic view of an X-ray diffraction experiment. The X-ray beam is produced by the source on the left and is conditioned by the optics before interacting with the crystal mounted on the goniometer. Scattered X-rays are detected on the CCD or imaging-plate detectors.

### 2.4 Data Processing

The observed intensity of a reflection is given by Darwin’s equation,

\[
I(hkl) = I_0 \cdot r_e^2 \cdot \left( \frac{\lambda^2}{\omega} \right) \cdot \{V_x / V^2\} \cdot L \cdot P \cdot A \cdot |F_{hkl}|^2
\]

where \(I(hkl)\) is the measured intensity on the detector, \(I_0\) is the incident beam intensity, \(V_x\) is the volume of the crystal illuminated in the X-rays, \(V\) is the volume of the crystal unit cell, \(r_e\) is the classical electron radius (2.818 x 10^{-15} m) of an electron, \(\lambda\) is the X-ray wavelength, \(\omega\) is the angular velocity of the crystal, \(L\) is Lorentz correction, \(P\) is polarization correction and \(A\) is the absorption correction [245-248]. The analysis and reduction of a single crystal raw diffraction data consists of seven major steps.

1) Visualization and preliminary analysis of the original, unprocessed detector data.
2) Indexing of the diffraction pattern, i.e., to determine the indices \(h, k, l\) of each reflection, as well as the parameters that define the size and shape of the crystal unit cell, and its orientation on the diffractometer.
3) Refinement of the crystal and detector parameters.
4) Integration of the diffraction maxima.
5) Apply geometric corrections and finding the relative scale factors between measurements.
6) Precise refinement of crystal parameters using whole data set.
7) Merging and statistical analyses of the measurements related by space group symmetry, putting all the reflections from all of the images on the same relative numeric scale and then merge and average them to produce a unique reflection list with associated intensities and realistic error estimation.

Several computer programs are available to perform all these steps. Widely used data reduction program packages are XDS [249-254], MOSFLM [255-257], DPS [258-261], d*TREK [262, 263] and HKL [241].

For the work presented in this thesis, the XDS software suite was used for data reduction/processing. The XDS software suite consists of the following programs: XDS for indexing [241, 251, 258, 264-273] and integration and XSCALE for scaling and merging data [253, 274-281].

A reduced cell is extracted from the observed diffraction pattern and processing of the data images continues to completion as if the crystal were triclinic. The reflection indices then refer to the reduced cell and reindexed once the space group is known. For all space groups, the required reindexing transformation is linear and involves only whole numbers [282]. Automatic space-group assignment is carried out in two steps once integrated intensities of all reflections are available [253]. Firstly, the Bravais lattices are identified that are compatible with the reduced cell derived from the observed diffraction pattern. In the second step, all enantiomorphous space groups compatible with the observed lattice symmetry are rated by a redundancy-independent R factor [283, 284]. The group is selected that explains all integrated intensities in the data set at an acceptable R factor requiring a minimum number of unique reflections.

The quality of the data is judged by the parameters, R-merged, I/σ (I) and completeness, at automatically determined resolution shells.

\[
R\text{-merge} = \sum_{hkl} \sum_{i} I_{i}(hkl) - <I(hkl)> \sum_{hkl} \sum_{i} I_{i}(hkl) \]  

The summation is over all observations, and the values of R are typically 0.05 - 0.12 for reasonably well-measured data. The criteria for the inclusion of the highest resolution shell should roughly
satisfy all or at least a pair of the following conditions: $I(hkl) / \sigma \{I(hkl)\} \geq 2.0$, completeness $\geq 75\%$ and $R$-merged $\leq 50\%$.

### 2.5 Phase solution

Once the diffraction data has been obtained, the next step is to solve for the phases of the reflections.

The protein structure solution scheme is given in Figure 15 and methods used are given in Table 3.

<table>
<thead>
<tr>
<th>Table 3: Methods used in phase solution.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
</tr>
<tr>
<td>----------------------------------------</td>
</tr>
<tr>
<td>Direct methods</td>
</tr>
<tr>
<td>Molecular replacement</td>
</tr>
<tr>
<td>Isomorphous replacement</td>
</tr>
<tr>
<td>Anomalous scattering</td>
</tr>
<tr>
<td>Density modification (phase improvement)</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

![Figure 15: The structure solution protocol.](image)

The methods such as Molecular Replacement (MR) [285-297, 300-305], Multiple Isomorphous Replacement (MIR) [306-309], Multiple Wavelength Anomalous Diffraction (MAD) and Single
Wavelength Anomalous Diffraction (SAD) [310-319] are used for this purpose. In this thesis, Molecular Replacement method is used and will be described next, followed by the other methods.

2.5.1 **Molecular replacement (MR) method**

Molecular replacement is fundamentally a search method of solving crystal structures when a suitable related model is available [285-297]. The model structure may correspond to a crystal structure or a solution structure of a homologue of the unknown structure. The required sequence identity should be > 25%. MR enables the solution of the crystallographic phase problem by providing initial estimates of the phases of the new structure from a previously known structure. MR is a good phasing method because it involves the collection of a single data set from a single crystal and has the advantage of minimizing the effects of radiation damage. With the success of structural genomics and proteomics projects, the number of available structures will increase rapidly, leading more harness of MR method. The pre-requisites for MR method are: a) an observed diffraction pattern (intensities) of the unknown target structure, and b) the atomic coordinates of a homologous probe protein structure.

The MR task involves positioning the probe within the unit cell of the target crystal in such a way that the theoretical diffraction pattern that would result from this model closely matches the experimental one. If only one molecule is present in the asymmetric unit, then six parameters (three rotational and three translational), which fully describe how the probe is placed in the unit cell, would have to be determined. In principle, one could do a search on these six parameters to determine the position of the probe that gives the best agreement between observed and calculated structure factor. However, such a search would be computationally too demanding. From the theoretical analysis of the properties of the Patterson function [298, 299] it became obvious that such a six-parameter search could be reduced to two three-dimensional problems. The first is the determination of the correct orientation of the probe, and the second is the determination of the position of the correctly oriented molecule within the unit cell. Therefore, the main aim of the MR method is to find these two operators to solve the rotation and the translation functions. The searches can be by using either match in the Patterson function or through likelihood functions. There have been several advances in the MR technique which extend its scope [290-297, 320, 321].
2.5.1.1 The Rotation Function

Patterson function is used to obtain the rotation function [286]. Patterson function \( P(\bar{x}) \) is the Fourier transform of \( |\tilde{F}(hkl)|^2 \) and can be calculated using the observed structure factor amplitudes without knowing the phases. If the Patterson function is calculated from the structure factor amplitudes of the crystal, \( P_1(\bar{x}) \) and that calculated from the model, rotated by the rotation operator \( \tilde{R} \), \( P_2(\tilde{R}\bar{x}) \), then the rotation function is given by,

\[
RF(\tilde{R}) = \int_U P_1(\bar{x})P_2(\tilde{R}\bar{x})dV
\]

where the integral is over a volume \( U \), and large values are obtained when there is a maximum overlap between \( P_1(\bar{x}) \) and \( P_2(\tilde{R}\bar{x}) \). Such a function describing the overlap, calculated in reciprocal space, has been first presented by Rossmann and Blow [286, 289]. A maximum in this rotation function \( RF(\tilde{R}) \) indicates a potentially correct orientation of the search probe in the target cell. However, the problem with the Rossman-Blow function was large computation time required and lack of precision owing to truncation errors in the interference function obtained from the integrals in equation 12. This is modified to a new form and was described as the fast rotation function by Crowther [300] which is based on spherical coordinate system and is computationally faster.

2.5.1.2 The Translation Function

Once the orientation of a probe molecule is known from the rotation function, the actual position of the molecule in the unit cell has to be determined. In the translation function the correlation between the observed intensities and the Patterson cross-vectors of the symmetry-related molecules of the probe is calculated as the molecule is moved within the cell [286, 322]. When the molecule is correctly positioned, the translation function should have peaks at values corresponding to the translation vectors between the symmetry related molecules. The combination of a three-dimensional translation function with a packing analysis [301] gives more realistic and accurate results. The translation is also determined from the correlation coefficient between the observed and calculated amplitudes squared, as a function of model translation. The Patterson correlation (PC) score is,
The goal of PC refinement is to improve the overall orientation of the search model and is carried out for rigid bodies such as domains, subdomains or secondary-structure elements. The target function of PC refinement is typically defined as the standard linear correlation between observed and calculated squared normalized structure-factor amplitudes ($|E|^2$). By improving the accuracy of the search model for the correct angular orientation, PC refinement improves the discrimination between correct and incorrect orientations and therefore enables the location of the correct peak in a noisy rotation function.

2.5.1.3 Maximum Likelihood based MR method

The likelihood of the model given the data is defined as the probability of the data given the model. Maximum likelihood (ML) model is that model which explains the measured data with highest probability [323]. The likelihood of the model given the data is defined as the probability of the data given the model. Where the data have independent probability distributions, the joint probability of the data given the model is the product of the individual distributions. In crystallography, the data are the individual reflection intensities which are not strictly independent. Systematic variation in intensities and correlations present due to non-crystallographic symmetry hinders estimation of joint probability as a simple product. Nevertheless, the assumption of independence and anisotropy corrections are necessary to make the problem tractable. The log of the likelihood function (LL) is used to give a reasonable numeric value compared to the product of small probabilities. In the present thesis Phaser is used for MR solutions. In the ML functions in Phaser, the structure-factor amplitudes (Fs), or normalized structure-factor amplitudes (Es) are used [324]. Integrating out the unknown phases gives the Rice distribution or Sim [325] distribution from a two-dimensional Gaussian distribution. The structure factors have four probability distributions: 1) for an acentric reflection, the probability distribution is (Rice distribution) dependent on $F_O$ with its expectation value, $<F_O>$ centered on $<F_O>$, 2) for a centric reflection, the probability distribution is a Woolfson distribution, 3) for an acentric reflection when the atoms are totally random, i.e. $<F_O> = 0$, the Rice becomes Wilson distribution and, 4) for a centric reflection and random atoms Woolfson becomes centric
Wilson distribution. In MR, the expection values \(<F_0>\) of the structure factors, using the set of free reflections, are calculated from the model, \(F_C\) multiplied by the fraction of correct structure, \(D (0 > D > 1)\). These probability distributions are the basis for all the maximum likelihood functions used in Phaser. In a typical run of the program, the fast rotation and translation functions are used to identify potential solutions, which are then rescored using the full likelihood functions to determine the best solution. Packing criteria are a very powerful constraint on the translation function. It has been shown that judicious editing can make a significant difference in the quality of a distant model. Any available phase information, be it from experimental phasing (i.e. multi-wavelength anomalous diffraction (MAD), single-wavelength anomalous diffraction (SAD), Multiple Isomorphous Replacement (MIR), etc.) or a partial molecular replacement solution, can be used in solving molecular replacement problems through calculations of phased-rotation and phased-translation functions. In models having poor sequence identity, an ensemble (e.g. ensemble of NMR models or homology models) may be provided in Phaser. The best ensembles are generally derived using alignment protocols, such as FFAS protocol [326]. Automated Molecular Replacement in Phaser combines the anisotropy correction, likelihood enhanced fast rotation function, likelihood enhanced fast translation function, packing and refinement modes for multiple search models and a set of possible space groups to automatically solve a structure by molecular replacement.

For molecular replacement, software suites AMORE [302, 303] and MOLREP [305] use fast rotation and translation functions, but both lack likelihood-based scoring functions. In MOLREP the search for multiple copies of a molecule in the unit cell and incorporation of the packing function provides better contrast indentifying correct solutions. Another approach in MR that is implemented in the program CNS [290, 292] is based on two concepts- i) the direct rotation search and ii) Patterson correlation (PC) refinement. In the traditional rotation search, the computationally expensive structure-factor calculation is carried out only once to obtain a model Patterson map, which is then rotated and superimposed with the observed Patterson map. In the direct rotation search, however, the model is rotated directly and a structure-factor calculation is carried out for each sampled angular orientation. This has the advantage of avoiding approximations such as interpolations, but has the disadvantage of being computationally expensive. The intervening step between the rotation search and the translation search in XPLOR/CNS is the PC refinement [290].
Programs to perform six dimensional searches similar to conventional MR are implemented in programs such as ‘Queens of Spades’, ‘EPMR’ and ‘SOMoRe’ [294, 295, 327].

2.5.2 Multiple isomorphous replacement (MIR)

This method of phasing involves the isomorphous attachment of heavy atoms to the protein molecules in the crystal [306, 307]. Since different atoms contribute to the scattered intensity in proportion to the square of the number of electrons they contain, a heavy atom like U, Hg, I, As, W, etc. which contain many electrons than atoms typically present in proteins, will contribute much more to the diffracted intensities. As a result, the change in intensity from the addition of one heavy atom to a typical protein is easily measured. If there are two crystals, one containing just the protein (native crystal) and other containing in addition bound heavy atoms (derivative crystal); one can measure diffraction data from both. The differences in scattered intensities will largely reflect the scattering contribution of the heavy atoms, and these differences can be used to compute a Patterson map [308].

Because there are only a few heavy atoms, such a Patterson map will be relatively simple and easy to deconvolute (Alternatively, direct methods can be applied to the intensity differences). Once the location of the heavy atoms in the crystal is known, their contribution to the structure factors can be calculated. Here the assumption is made that the heavy atom doesn't change the rest of the protein structure and scattering from the protein atoms is unchanged by the addition of heavy atoms. Thus, the structure factor for the derivative crystal \( \tilde{F}_{PH} \) is equal to the sum of the protein structure factor \( \tilde{F}_P \) and the heavy atom structure factor \( \tilde{F}_H \),

\[
\tilde{F}_{PH} = \tilde{F}_P + \tilde{F}_H
\]

In this equation only the lengths of \( \tilde{F}_{PH} \) and \( \tilde{F}_P \) are known, but \( \tilde{F}_H \) is known both in length and direction. Through Harker construction [309], it turns out that there are two ways to draw vector triangle that will satisfy above vector equation leading to two possible phases of \( \tilde{F}_P \). This twofold phase ambiguity can be removed by preparing a second derivative crystal with heavy atoms that bind at other sites. Only one phase choice of \( \tilde{F}_P \) will be consistent with all three observations. Hence, in principle, at least two heavy atom derivatives are required to get the phase information.
2.5.3 Multiple wavelength anomalous diffraction (MAD)

With the advent of powerful synchrotron sources, this method exploits the fact that, at certain wavelengths, tightly bound electrons in an atom do not behave like free electrons in the scattering process [205, 310]. This causes the anomalous scattering, an effect that depends on the wavelength and is generally stronger for heavy atoms [205, 311, 312]. The consequence of anomalous scattering is that the intensities of certain otherwise equivalent reflections called Friedel and Bijvoet mates [311-318], are no longer equal, and this difference in intensities can be used for phase angle determination. This method facilitates phase determination from one single crystal and the earliest experiments showing the practical feasibility of this approach were carried out by Ramaseshan et al. in 1957 [205]. Most often Se-atom is introduced in the protein by growing the microorganisms on S-methionine substrate instead of the normal methionine containing substrate [319, 328]. The single-wavelength anomalous dispersion (SAD) can be used as a phasing method [329-333]. The advantage is that a single dataset is required but the statistical accuracy has to be increased by collecting a highly redundant intensity dataset. Another advantage is that in-house laboratory sources can be used to collect a sulphur-SAD dataset utilizing the intrinsic sulphurs of proteins [332, 333].

2.5.4 Direct methods

Direct methods are based on the positivity and atomicity of electron density that leads to phase relationships between the (normalized) structure factors. For developing direct methods for phasing Hauptmann and Karle shared the 1985 Nobel Prize in Chemistry [334-339]. The direct method attempts to calculate the phases directly from the magnitudes of normalized structure factors $|\tilde{F}_h|$. The method relies on the fact that certain linear combinations of phases are structure invariants, which do not depend on the choice of origin. The most useful structure invariants are triplet invariants $\Phi_{hk} = \phi_h + \phi_{-k} + \phi_{-h+k}$ where $h$ and $k$ are two arbitrary indices of the reciprocal lattice and $\phi_h$ is the phase of a structure factor $\tilde{F}_h$. The triplet relation shows how the phases of three reflections are related. Therefore, knowing the phases of two reflections and the structure invariant $\Phi_{hk}$ allows one to derive the phase of a third. Assuming a random distribution of identical atoms, the probability distribution of the invariant, $P(\Phi_{hk})$, is given by,
\[ P(\Phi_{hk}) = \left(\frac{1}{L}\right) \exp\left(G_{hk} \cos \Phi_{hk}\right) \]  

(15)

where \( L \) is a normalization constant and \( G_{hk} = \left(\frac{2}{\sqrt{N}}\right)|E_h E_k E_{h-k}| \); \( N \) being the number of atoms in the unit cell. This distribution is maximal when \( \Phi = 0 \) and decreases as \( \Phi \) deviates further from 0. Also, \( P(\Phi_{hk} = 0) \) increases as \( |E_h E_k E_{h-k}| \) increases and \( N \) decreases. Each reflection is involved in multiple invariants having multiple probability distributions. The tangent formula [335],

\[
\tan \phi_h = \frac{\sum_{j=1}^{r} G_{hk} \sin(\phi_{j} + \phi_{h-k})}{\sum_{j=1}^{r} G_{hk} \cos(\phi_{j} + \phi_{h-k})}
\]  

(16)

combines these distributions and assigns the most probable phase of \( h \) given other phase angles. Most classical direct methods use the probability distributions or the tangent formula to generate phases with high overall probability. As the number of atoms \( (N) \) increases, the probability distribution flattens. One of the reasons the classical direct method fails for molecules with more than \( \sim 200 \) atoms in the asymmetric unit, is that it becomes much harder to predict the phases when \( P(\Phi_{hk}) \) is almost the same for all values of \( \Phi_{hk} \).

The direct methods are also being utilized increasingly now in combination with anomalous-dispersion methods and isomorphous replacement methods to locate the position of heavy or anomalous atoms from a very small signal using direct methods programs, SnB [340-344] and SHELXD [345, 346]. Several programs, mostly based on maximum-likelihood principles, are available for carrying out heavy-atom refinement and phasing. These include: CNS [347], MLPHARE [348], SHARP [349], SOLVE [350], and HySS [351].

2.5.5 Other method of phasing

A new phasing method [352-354], called VLD (Vive la Difference), is based on difference and hybrid Fourier syntheses [355-357]. It is claimed that this new method can obtain the correct difference structure even when the model is completely uncorrelated or is completely random from the true structure. VLD has been implemented in the program IL MILIONE [358]. It can also use a starting seed model, found by other phasing methods (e.g., Patterson techniques, MR, SAD, MAD, SIR-MIR), and extend and refine the phases. This phasing algorithm is suitable for structural solution in Structural Proteomics projects.
2.6 Density modification

Once the phase information is available and a preliminary model of the protein is known, the next step is the calculation of the electron density map. The quality of the phases determines the quality of the resulting map. There are methods to improve the initial electron density maps, and these are referred to as density modification methods [359-361]. Density modification works by incorporating features that are expected to appear in a correct map, such as the similarity of regions related by non-crystallographic symmetry [362], flatness or disorder of the solvent region [329, 363, 364] and the similarity of the density-map histogram to histograms of macromolecules whose structures have been determined earlier [365]. Histogram matching is a robust statistical density modification technique which is done along with solvent flattening. The average features of electron density of known proteins are similar, as they are composed of twenty amino acids, and this is utilized in obtaining the frequency distribution of the electron density values. These distributions are a function of electron density, and are independent of protein at a given resolution. The plots of frequency distributions of the electron density levels calculated at grid points as a function of electron density are scaled to resemble a standard plot. Solvent flattening and averaging improve the phases of low resolution reflections and when combined with histogram matching, the phases of higher resolution reflections improve considerably [360].

2.7 Electron density map interpretation

Once the initial electron density maps are improved, the next step is the interpretation of these maps, which is nowadays achieved with the help of a broad spectrum of programs and powerful computers. Interactive computer graphics programs O [366] and Coot [367] were used to build the model into the calculated $2mFo-DFc$ electron density maps [368]. The model is usually built by interpreting simultaneously two types of electron density maps, the $mFo-DFc$ [369] and the $2mFo-DFc$ map contoured at 2 $\sigma$ and 1 $\sigma$ levels respectively. The $mFo-DFc$ map contains positive peaks where density is not accounted for by the model used to calculate the $Fc$, and hence is useful to locate missing or wrongly placed atoms. The $2mFo-DFc$ map is the sum of a $Fo$ map and $mFo-DFc$ map, and contains information from both the $Fo$ map and the difference $mFo-DFc$ map. In the simulated annealed (SA) omit electron density map calculation, portion/portions of the model in the region of
interest is/are left out and the rest of the model is used in phase calculations reducing model bias [370, 371].

2.8 Crystallographic refinement

The aim of refinement is optimization of the atomic model to simultaneously fit both observed diffraction data and a priori chemical information. The refinement of a macromolecular structure is a difficult optimization problem because of poor parameters to observations ratio and because of correlations between the coordinate of each atom to all the other (nearby/bonded) atoms [372]. The target functions are functions of many parameters and their errors [373, 374]. The large number of adjustable parameters results in the multiple minima problem — the target function contains many local minima in addition to the desired global minimum. The challenges of crystallographic refinement arise not only from this high dimensionality of the parameter space, but also from the poorer quality and/or lower resolution of experimental phases which guide correct shifts to atomic parameters.

The refinement programs differ in the nature of the target function(s) and the method(s) by which the target function is minimized. In energy-based methods, such as CNS [370] and PHENIX [375] the target function is expressed as energy and the weights of the restraints are estimates of the energy penalty associated with deviations from the target values. Other refinement programs are implemented in REFMAC [376], TNT [377], and SHELX [378]. The force constants and stereochemical parameters are derived from parameter sets of Engh and Huber [379] and Parkinson et. al. [380], derived from the high-resolution small molecule crystal structures. All the structures reported in the present thesis were refined using maximum likelihood target function and simulated annealing optimization to overcome multiple minima problem [372].

2.8.1 Target functions

Crystallographic refinement is a search for the global minimum of the target,

\[ E = E_{chem} + w_{xray} E_{xray} \]  

(17)

where \( E \) is a function of the parameters of an atomic model. \( E_{chem} \) comprises empirical information about chemical interactions; it is a function of all atomic positions, describing covalent (bond lengths, bond angles, torsion angles, chiral centers and planarity of aromatic rings) and non-bonded
(intramolecular as well as intermolecular and symmetry related) interactions [381, 382]. \( E_{\text{ray}} \) is related to the difference between the observed (\( F_\text{o} \)) and calculated data (\( F_\text{c} \)) and \( w_{\text{ray}} \) is a weight appropriately chosen to balance the gradients (with respect to atomic parameters) arising from the two terms.

### 2.8.2 Maximum Likelihood refinement targets

The most commonly used target function (\( E^{\text{LSQ}} \)) for macromolecular refinement employs the least-squares residual for the diffraction data [383],

\[
E^{\text{LSQ}} = E_{\text{restraints}} + w_a \sum_{hkl} \left( |F_\text{o}| - k |F_\text{c}| \right)^2
\]

where \( |F_\text{o}| \) and \( |F_\text{c}| \) are the observed and calculated structure-factor amplitudes, \( k \) is a relative scale factor, \( w_a \) is a weight, and \( E_{\text{restraints}} \) are geometric (bond length, bond angle, and atomic repulsion) restraints. A decrease of this function can sometimes be due to accumulation of systematic errors in the model without improvement or even a worsening of the model. The underlying reason can be found in the fact that the least-squares residual does not account for the effects of phase errors in the calculated structure factors, so it is poorly justified when the model is far away from the correct answer or incomplete. A more appropriate target for macromolecular refinement can be obtained through a maximum likelihood formulation [323, 384-389]. The goal of the maximum likelihood method is to determine the probability of making a set of measurements, given the model, and estimates of its errors and of errors in the measured intensities [372]. The effects of model errors (misplaced atoms and missing atoms) on the calculated structure factors are first quantified with \( \sigma_a \) values, which correspond roughly to the fraction of each structure factor that is expected to be correct [371]. To achieve an improvement over the least-squares residual (Equation 19), cross-validation [390-393] was used for the computation of \( \sigma_a \), necessitating its calculation with a randomly selected test set of diffraction data that was never included in the refinement process. The cross-validated \( \sigma_a \) values (\( \sigma_a^{\text{cv}} \)) are then used to compute the expected value of \( \langle |F_\text{o}| \rangle^{\text{cv}} \). \( \langle |F_\text{c}| \rangle^{\text{cv}} \) and \( \sigma_{\text{ME}^{\text{cv}}}^2 \) can be readily incorporated into a maximum likelihood target function,

\[
E^{\text{ML}} = E_{\text{restraints}} + w_a \sum_{hkl \in \text{working set}} \left( \frac{1}{\sigma_{\text{ME}^{\text{cv}}}^2} \right) \left( |F_\text{o}| - \langle |F_\text{c}| \rangle^{\text{cv}} \right)^2
\]
The least-squares residual is a limiting case of the maximum likelihood theory and is only justified if the model is nearly complete e.g. rigid-body refinement [394]. For many structures, some initial experimental phase information is available from either MIR or MAD methods. These phases represent additional observations that can be incorporated in the refinement target. The addition of experimental phase information in the target function greatly improves the results of refinement [388].

2.8.3 Minimization method - Simulated Annealing

The suitable target function has to be minimized. In protein structure refinement, two minimization methods are used- a) gradient descent and b) simulated annealing. Crystallographic target function has multiple minima and it is difficult to cross the barrier between the minima by the gradient descent method. Simulated annealing method [395-399] can overcome the multiple minima having greater radius of convergence, thereby reducing the manual intervention required during refinement. By defining the target $E$ (equations 18 and 19) to be the equivalent of the potential energy of the system, one can simulate annealing process. The likelihood of uphill motion is determined by a control parameter referred to as temperature. The higher the temperature, the more likely it is that simulated annealing will overcome barriers. The two most commonly used protocols are linear slow cooling or constant temperature but better model is obtained with slow cooling [397]. Many examples have shown that simulated annealing refinement, starting from initial models obtained by standard crystallographic techniques, produces significantly better final models compared to those produced by conjugate-gradient minimization method [400]. The parameter space can be reduced drastically for proteins by employing simulated annealed torsion angle molecular dynamics [401, 402].

In the present thesis, CNS, CCP4 and Phenix were used for refinement. The structures were refined in CNS using standard simulated annealing protocols and the amplitude based maximum likelihood target function [400, 402]. A total of 5% of randomly selected reflections were set aside for cross validation [391]. All reflections in the respective resolution ranges were included in the refinement. In the initial stages of SA refinement and during calculation of SA omit maps, the model was heated to a temperature of 2000 °C, and then annealed at a cooling rate of 25 °C per iteration. Initial anisotropic B-factor and bulk solvent corrections were applied [403]. The relative weighting between geometric and X-ray terms in the target function was determined automatically in CNS. The
parameters and the minimization methods used in various common refinement programs are documented [404].

2.8.4 Multi-start refinement

Multi-start refinement is multiple SA refinements starting from a single model [405]. Some of the models resulting from multi-start refinement may give lower free $R$-value. Each model coming from a multi-start refinement fits the data slightly differently and more variance indicates intrinsic flexibility within the molecule. Also the regions in the starting model that contain significant errors or poor electron density show increased variability after multi-start refinement and a visual inspection of the ensemble of models produced can be helpful in identifying these incorrectly modeled regions.

In order to better identify the correct conformation, structure factors from each of the models can be averaged [405]. This averaging tends to reduce the effect of local errors (noise), which are presumably different for each member of the family. The average structure factors produce phases that contain less model bias than phases computed from a single model. It also produces better estimates of errors in the model for maximum likelihood targets and $\sigma_A$-weighted electron density maps because $F_c$ is used in the computation of these parameters. Multi-start refinement followed by structure-factor averaging is useful in situations in which there is significant noise, namely when the data to parameter ratio is very low. The program wARP [406] makes efficient use of structure-factor averaging in the context of phase improvement and automated model completion.

2.8.5 TLS Refinement

TLS is rigid body collective displacements of whole molecules, domains and secondary structure elements. TLS stands for translation, liberation and screw motions comprising of 20 TLS parameters per group [407-409]. In this refinement, motions of groups of atoms are utilized, instead of individual atoms, which reduce the number of parameters drastically. A protein chain is divided into few groups depending upon domains, flexible and rigid groups [408]. After fitting the model, PHENIX [375] and CCP4 [376] were used for TLS refinement on the protein model.

2.8.6 R-factor and cross-validation

A widely used indicator to represent the correctness of the model structure is the R-factor or the disagreement factor. It is defined as,
\[ R = \frac{\sum_{hkl} |F_{\text{obs}}| - kF_{\text{calc}}|}{\sum_{hkl} |F_{\text{obs}}|} \times 100\% \]

For a well-refined protein structure it is generally below 20%. A low R-factor means that the model agrees with the diffraction data more closely. However there is always a danger of overfitting the diffraction data. To reduce this danger the concept of cross-validation in the form of the free R-value has been introduced [393]. Cross-validation also produces more realistic coordinate error estimates based on the Luzzati or \( \sigma_A \) methods [410, 371]. For cross-validation, the diffraction data are divided into two sets—a large working set (typically comprising 95% of the data) and a complementary test set (comprising the remaining 5%). The diffraction data in the working set are used in the normal crystallographic refinement process, whereas the test set data are not. The cross-validated R-value (or \( R_{\text{free}} \)) computed by equation 20 using the reflections in the test set is a better indicator of model quality. It provides a more objective guide during the model building and refinement process than the conventional R-value that is computed by above equation using reflections in the working set. \( R_{\text{free}} \) also indicates whether the introduction of additional parameters (e.g. water molecules, the relaxation of non-crystallographic symmetry restraints or multi-conformer models) improves the quality of the model or, instead, increases overfitting.

Water molecules are added manually by examining environment around electron densities that were present in both \( m\text{Fo-DFc} \) and \( 2m\text{Fo-DFc} \) maps. A composite omit map, as implemented in CNS, is calculated to enable unbiased interpretation of the electron density map. Composite omit maps are calculated by leaving out 3% of the amino acid residues at a time. All superpositions of the structures are carried out using the softwares \( O \) [366] and \( Coot \) [367].

2.9 Validation of the model

The last step in protein structure determination is the validation of the model. The crystallographic R-value is an indicator of the model quality, but it has been shown that \( R_{\text{free}} \) is more reliable. The computer program \( PROCHECK \) [411] has been used to check the stereochemical and geometrical parameters of the model by comparing them with ideal values obtained from a database of well refined high-resolution protein structures in the Protein Data Bank (PDB) [412]. The checks performed are on covalent geometry, planarity, dihedral angles, chirality, non-bonded interactions,
main chain hydrogen bond, disulfide bonds, stereochemical parameters and residue-by-residue comparisons. It checks the main-chain bond lengths and angles against the ideal values given by Engh and Huber analysis [379] of small molecule structures in Cambridge Structure Database [413]. The computer program Coot [367] uses interactive graphics outputs of various plots and a residue-by-residue listing, providing an assessment of the overall quality of the structure and solvents highlighting regions which may need further investigation. The Ramachandran plot [414] of the structure is given by both the programs. Most of the residues in a protein model should lie within the allowed regions of the Ramachandran plot. CNS is also used to analyze geometrical parameters and list rms deviations in bond lengths, bond angles, dihedral angles, short contacts between atoms etc. All the structures were validated by using ADIT prior to PDB submission [412]. The figures were made using softwares Pymol [415] and Chimera [416]. Real-space correlation coefficients of the final refined models provide the accuracy of the positions of the atoms within the positive electron density maxima [417]. More precisely the SA omit electron density map is correlated to the positions of the refined atoms and are used in real-space correlation coefficients calculations. Automatic fitting into the electron density, building the atomic model and refinement are incorporated in current programs viz. Coot, Phenix-Autobuild, ARP/wARP, etc. [367, 375, 417-420].

The atomic coordinate errors can be found by inversion of least-squares full matrices provided an atomic or ultra-high resolution protein structure is present [378]. But when normal resolution structures are available, the coordinate errors can be estimated by using Cruickshank’s empirical formula (Equation 21) for an atom with \( B = B_i \) [421].

\[
\sigma_i(r, B_{\text{avg}}) = k \left( \frac{N_i}{p} \right)^{1/2} \cdot \left[ g(B_i) / g(B_{\text{avg}}) \right] \cdot C^{-1/3} \cdot R \cdot d_{\text{min}}
\]

where \( k = 1.0, N_i = \sum Z_j^2 / Z_i^2, p \) is difference between number of observations and parameters, \( B_{\text{avg}} \) is the average B for fully occupied sites, \( C \) is the fractional completeness of the data to \( d_{\text{min}} \), and \( R \) is refinement R-factor. \( g(B) = 1 + a_1B + a_2B^2 \) is an empirical function to allow for the dependence of \( \sigma_i(r) \) on \( B \). The parameters \( a_1 \) and \( a_2 \) depend upon the structure [422]. The \( \sigma(r) \) is called diffraction-component precision index (DPI). This formula may be used to estimate the errors in the hydrogen bonding distances in a refined protein structure. In all the figures in subsequent chapters, 3.4 Å was used as the hydrogen bond length cutoff. The next chapter shall provide specific experimental details of HIV-1 protease preparation, data collection and structure of complex with a type-2 substrate.