Chapter 4
Structure Determination, Model Building,
Refinement and Validation of
Pab-NTD and Pab-NTD-L-serine complex
Chapter 4-Structure Determination, Model Building, Refinement and Validation of Pab-NTD and Pab-NTD- L-serine complex

4.1 Introduction

A macromolecular structure is determined from X-ray diffraction pattern, by calculating the electron density distribution inside the crystal. The diffraction pattern of a crystal is related to the electron density through a mathematical operation called the Fourier transform. Computation of the inverse Fourier transform of the diffraction will give the electron density $\rho(x, y, z)$, which is given by following equation:

$$\rho(x, y, z) = \frac{1}{V} \sum_{hkl} F(h, k, l) \exp[-2\pi i (hx + ky + lz) + i\alpha(h, k, l)]$$

where $F(h, k, l)$ is the structure factor for the corresponding reflection $(h, k, l)$ and $\alpha(h, k, l)$ is the phase of that reflection. Thus, in order to calculate the electron density map, we must have accurate information of both structure factor amplitude $F(h, k, l)$ as well as the phase $\alpha(h, k, l)$. In a typical X-ray diffraction experiment, a count of X-ray photons on each spot is measured, which gives the intensity that is proportional to the square of the amplitude (peak height) of the diffracted wave. However, there is no practical way of measuring the relative phase angles for different diffracted spots experimentally and thus the phase information is lost. Hence, in the absence of phase information, electron density cannot be calculated and this is referred to as "The Phase Problem" in crystallography. Apart from growing good diffraction quality crystals, this is one of the major bottlenecks in structure determination. Since phase of a reflection could not be measured directly, it can be deduced indirectly by using one of the three commonly used techniques: Molecular Replacement (MR), Multiple wavelength Anomalous Diffraction (MAD) and Multiple Isomorphous Replacement (MIR) method.

Molecular Replacement is the most common technique used to solve the structure of a protein belonging to the same structural family. It utilizes the information coming from the known structure of the homologous protein to determine phases of an unknown structure. Multiple isomorphous replacement technique is used for phase determination by incorporating heavy atoms in the native crystals. This causes change in the diffraction intensity of the derivatized crystals. The difference in the intensities of native and derivatized crystal is used to determine the protein phases. MAD is a conceptually similar technique where the protein crystal is incorporated with an anomalous scatterer, mostly selenomethionine, and is used for phase calculation. In this case, the difference in the
Chapter 4-Structure Determination, Model Building, Refinement and Validation of Pab-NTD and Pab-NTD- L-serine complex

intensities between Friedel or Bijvoet pairs of reflections, occurring due to the presence of anomalous scattering atoms, is used for phase calculation.

Pab-NTD, as discussed earlier, is a protein with a unique sequence and hence the structure solution by using MR was not attempted. Out of 183 amino acid residues, Pab-NTD contained only one methionine and that too at the N-terminus. A frequency of one methionone per 80 residues has been estimated to obtain measurable anomalous signals. When compared to this, the frequency of methionine in Pab-NTD is too less and therefore the possibility of solving structure by MAD was also ruled out. Hence, MIR method was used to solve the structure. Pab-NTD-L-serine structure was solved by MR method, using the wild type Pab-NTD structure as a model.

4.1.1 Isomorphous Phasing

Isomorphous replacement phasing technique is based on the fact that the structure factor $F_{hkl}$ for a certain reflection $hkl$ is a simple summation of all individual atomic scattering contributions $f_j$

$$F_{hkl} = \sum_j f_j \exp [2\pi i (hx_j + ky_j + lz_j)]$$

Different atoms contribute to the scattered intensity in proportion to the square of the number of electrons they contain. For example, a uranium atom contains 15 times as many electrons as a carbon atom and therefore its contribution to the intensity will be to 225 fold to that of carbon atoms. Thus the introduction of a heavy atom will change the scattered intensity significantly. It is due to the fact that the "heavy" atoms contribute disproportionately to the overall intensity. The contributions from the lighter atoms will tend to cancel out, because they will scatter with different phase angles. However, all of the electrons in a heavy atom will scatter essentially in phase with one another. If the heavy atom doesn't change the rest of the structure i.e. the crystal is isomorphously derivatized, the structure factor for the derivative crystal ($F_{PH}$) is equal to the sum of the protein structure factor ($F_p$) and the heavy atom structure factor ($F_H$), or

$$F_{PH} = F_p + F_H$$
Chapter 4-Structure Determination, Model Building, Refinement and Validation of Pab-NTD and Pab-NTD- L-serine complex

Based on the assumption that the derivative crystal is isomorphous, some information could be derived on the structure factor amplitudes for the heavy atom from the differences between the derivative and the native dataset. It is important to remember that structure factors are complex numbers with a ‘real’ (magnitude) and a ‘imaginary’ (phases) component and can be thought of as vectors (Figure 4.1). Unfortunately, in the absence of phase information we cannot apply the simple subtraction.

\[ F_H = F_{PH} - F_P \]

4.1.2 Phase Determination by Isomorphous Replacement

The objective of a phasing experiment is to derive the unknown phase \( \alpha_p \) of each protein reflection. During data collection the structure factor magnitudes of \( |F_p| \) (protein) and \( |F_{PH}| \) (derivative) are measured. The differences in scattered intensities will largely reflect the scattering contribution of the heavy atoms \( |F_H| \), and these differences can be used to compute a Patterson map. Because there are only a few heavy atoms, such a Patterson map will be relatively simple and easy to deconvolute and the positions of the heavy atoms could be assigned. This gives the information of structure factors \( F_H \) of the heavy atom in one step. This information in turn is used to assign the phases (signs) of \( |F_P| \) and \( |F_{PH}| \), respectively.
Figure 4.2 Harker diagram showing determination of phase in an acentric case. (a) The Harker construction for a single derivative where intersection sites give the two probable values of $\alpha_p$ leading to phase ambiguity. (b) Harker construction for the MIR phasing, with two derivatives, where the intersecting point for all the three circles give the value for the phases without phase ambiguity.
4.1.3 The Phase Ambiguity

Structure factors of the reflections, belonging to a reflecting plane (Bragg's plane) parallel to an even-fold symmetry axis are 'real'. Their 'imaginary' part (phase) have only one of two possible values differing by $180^\circ$ e.g. $0/180^\circ$ or $90/270^\circ$ (Figure 4.1a). Such reflections are called "centric reflections" and the respective reflection plane is called as "centric zone" for example $P2_1$ (unique b axis) has one centric zone: $h0l$; $P2_12_12$ has three: $hk0$, $h0l$ and $0kl$. Isomorphous phasing becomes really advantageous and accurate for such centric zone reflections as $\alpha_p$ can take only one of the two possible values. In such situations the signs for $|F_p|$ and $|F_{PH}|$ could be easily assigned according to the following rule,

$$|F_{PH}| > |F_p|; \text{ both have same sign as } F_H$$
$$|F_{PH}| < |F_p|; \text{ both have sign opposite to } F_H$$

except for the rare cross over situation when $|F_p|$ or $|F_{PH}|$ is less than $|F_H|$ and their signs may differ. For centric reflections, $|F_H|$ could be directly estimated as difference between $|F_p|$ and $|F_{PH}|$.

However, the situation is entirely different for a non-centric zone reflection, also called "acentric reflection", where $\alpha_p$ can take any value from $0^\circ$ to $360^\circ$. In general, there could be two possible symmetric solutions for $\alpha_H$, unless $\alpha_p = \alpha_H$ or $\alpha_H + 180^\circ$, in which case the acentric becomes similar to the centric one (Figure 4.1b). For acentric reflections, $|F_H|$ could be estimated as

$$|\Delta F|_{iso} = |F_{PH}| - |F_p|$$

$$\approx |F_H| \cos (\alpha_{PH} - \alpha_H)$$

Although, a single isomorphous derivative provides useful information, still the sign of $|F_p|$ cannot be assigned uniquely by this method. It is due to the possibility of two values of phase for a given reflection and the problem is termed as "Phase Ambiguity". This could be readily explained by constructing a Harker diagram (Figure 4.2), which is another way of depicting the phase determination method. From the X-ray diffraction, we know only the magnitudes $|F_{PH}|$ and $|F_p|$. Since their phases can take any value from $0^\circ$ – $360^\circ$, it can be represented in the complex plane as a circle of radius $|F_{PH}|$ and $|F_p|$, respectively. If we know both the magnitude and the phase of $F_H$, we can draw both
circles offset by vector $|F_{H}|$ and obtain two solutions for possible phase values for $F_P$ (Figure 4.2 a).

In order to eliminate the "Phase Ambiguity" a second derivative can be prepared. On repeating the procedure, we can now obtain a unique solution for $\alpha_P$, the phase of $F_P$, provided that both the heavy atoms are not occupying the same position. Thus 'Multiple' derivatives could be used to obtain a less ambiguous phase information and this method is called as Multiple Isomorphous Replacement phasing (MIR).
\( \alpha_p \) and is given as the difference between \( F_{PHobs} \) and \( F_{PHcalc} \). The expression for \( F_{PHcalc} \) could be written as,

\[
F_{PHcalc} = |F_{Pcalc} + F_{Hcalc}|
\]

**Figure 4.3** Lack of closure error. Due to the error in the estimation of phases, \( F_{PHobs} \) would not be equal to \( F_{PHcalc} \) and the triangle made by \( |F_P|, |F_H| \) and \( |F_{PH}| \) will not close perfectly leading to lack of closure error. However, in case of most accurate phase calculation this triangle will close perfectly.
In MAD method, due to the anomalous scattering by heavy atoms there will be difference in the intensities between Friedel or Bijvoet pairs of reflections for the derivatized crystals. Because of the violation of Friedel's law, there are three different observations, \(|F_F|, |F_{PH+}|, \text{ and } |F_{PH-}|\), where + and - refer to the Bijvoet mates. In this case, anomalous differences between Bijvoet pairs, \(|\Delta F|_{ano}\) is actually used for phase calculation,

\[
|\Delta F|_{ano} = |F_{PH+} - F_{PH-}|
\]

\[
\approx |F_H| \sin (\alpha_{PH} - \alpha_H)
\]

It can be noted that, the information coming from anomalous scattering and isomorphous phasing is complementary to each other. On combining \(|\Delta F|_{iso}\) and \(|\Delta F|_{ano}\), a better estimation for the phases could be made. This method is termed as Single Isomorphous Replacement phasing with Anomalous Scattering (SIRAS) or Multiple Isomorphous Replacement phasing with Anomalous Scattering (MIRAS) depending on the number of heavy atoms used for phasing. In the case of Pab-NTD, the structure was solved using the MIRAS method.

4.1.4 The MIRAS technique

Once the native and derivative datasets have been collected and processed, the following steps are involved in order to solve macromolecular structures by MIRAS method:

1. **Scaling** of each derivative dataset with the native dataset.
2. Location of the heavy atom sites by calculating difference Patterson maps.
3. Initial phase estimation.
4. Completion of heavy atom model by calculating difference Fourier to locate and check for other sites.
5. **Refinement** of the heavy-atom parameters.
6. Calculation of the most probable phase and the best phase.
7. Calculation of the electron-density map.

---

65
4.1.5 Patterson Search for Heavy Atom Sites

Patterson function is a phase less Fourier synthesis of diffraction intensities (Patterson, 1935). It gives a map containing peaks at the endpoints of vectors between all the pairs of atoms in a unit cell of the crystal and all the non-origin peaks in a Patterson map represent inter-atomic vectors. In MIRAS method, a heavy atom site could be located by deconvolution of a difference Patterson map. An isomorphous difference Patterson map could be created using $\triangle|F|^2_{iso}$ as a coefficient to locate heavy atoms. The heavy atom site, once located in this map, could be further compared with approximately equivalent anomalous difference Patterson map calculated using $\triangle|F|^2_{ano}$ as coefficients. This shall give an added confidence to the location of the first heavy atom site.

The Patterson search procedure works by obtaining the heavy atom site from the 'Harker' vectors (vectors between the symmetry related atoms in the unit cell) concentrated on 'Harker' sections, which are always perpendicular to the rotation or screw axis. Once the first site for a particular derivative is chosen from the Harker vectors, it is fixed as the origin for all other sites in that derivative. Second and subsequent sites could be searched by looking at both Harker vectors for the site currently being searched for, as well as cross vectors (vectors between sites not related by space-group symmetry) between the current site and all the previously determined sites.

4.1.6 Initial Phase Estimation and Lack of Closure Error ($\varepsilon$)

Once the coordinates for heavy atom sites are obtained by difference Patterson map, an initial estimation of protein phases ($\alpha_p$) is done by using following formula,

$$\alpha_p = \alpha_H + \cos^{-1} \left( \frac{(F_{PH}^2 - F_p^2 - F_H^2)}{2F_p F_H} \right)$$

In case of most accurate phase calculation, the values of $F_{PHobs}$ and $F_{PHcalc}$ will be same. As a result, in the Argand diagram for isomorphous replacement phasing triangle made by $|F_p|$, $|F_H|$ and $|F_{PH}|$ will close perfectly (figure 4.1b). However, due to erroneous estimation of protein phases, there will be difference between $F_{PHobs}$ and $F_{PHcalc}$. This leads to an error called as "Lack of Closure" error, where the triangle will not close perfectly (Figure 4.3). "Lack of Closure" is an estimation of errors for the calculation of
Figure 4.6 Location of heavy atom sites in the Patterson maps. (a) Isomorphous difference Patterson map (3.5 Å) for gadolinium chloride (Gd) with the arrows indicating the Patterson peak (3σ) for heavy atom vector (0.10, 0.23). (b) Anomalous difference Patterson map for gadolinium chloride with the Patterson peak (5σ) marked at the same location.
Chapter 4-Structure Determination, Model Building, Refinement and Validation of Pab-NTD and Pab-NTD- L-serine complex

4.1.7 The Most Probable Phase and The Best Phase

The informations obtained by difference Fourier map, calculated by using inaccurate estimated phase would also be inaccurate. Thus, in the subsequent steps heavy atom parameters are refined. Heavy atom parameters such as occupancy, positional coordinates, scale factor and thermal parameters are refined to get maximum convergence of $F_{\text{PHcalc}}$ to $F_{\text{PHobs}}$, thereby reducing "lack of closure" error in each cycle of refinement. By calculating difference Fourier, a few other heavy atom sites are also located and incorporated in the heavy atom model, which will be further used for accurate phase calculation. The protein phases are finally calculated in terms of their probabilities due to the errors in data, structure solution and above all rare isomorphism. This could be done by two methods: Blow and Crick method and Hendrickson and Lattman method.

4.1.7.1 Blow and Crick method

This method is based on the assumption that all errors lie in the magnitude of $F_{\text{PH}}$, which could be given by "Lack of Closure" ($\varepsilon$) (Blow & Crick, 1959). The values of $\varepsilon$ for all the values of $\alpha$ (0-2$\pi$) are calculated. Assuming a Gaussian distribution of errors, the probability that the value of $\alpha$ for a given $F_p$ is correct, is related to the "Lack of Closure" ($\varepsilon$) of the phase triangle, and is given by,

$$P\alpha = P_j(\varepsilon) = \exp\left[- \frac{\varepsilon_j^2(\alpha)}{2E_j^2}\right]$$

where $E^2$ is the mean square value of "Lack of Closure"($\varepsilon$). When several heavy atom derivatives are used simultaneously, a relative probability could be calculated as follows,

$$P\alpha = \prod_j P_j(\alpha) = \exp\left[- \sum_j \varepsilon_j^2(\alpha) / 2E_j^2\right]$$

The "best phase" lies at the centroid of the phase probability curve whereas the "most probable phase" corresponds to the maxima of the centroid (Figure 4.4). It is the best phase that is used for the generation of electron density map of the protein structure.
4.1.7.2 Hendrickson and Lattman method

The major drawback of Blow and Crick method is storage of all the data. Incorporation of information coming from a new heavy atom derivative requires recalculations of whole probability curve. Hendrickson and Lattman method overcomes this problem due to a more compact form of storage. This method lies on the assumption that “lack of closure” (ε) lies in the derivative intensity $F_{PH}^2$ (Hendrickson & Lattman, 1970). Assuming a Gaussian distribution of the errors, the probability function can be cast in the form

$$P(\alpha) = \exp (A \cos(\alpha) + B \sin(\alpha) + C \cos(2\alpha) + D \sin(2\alpha))$$

where $A$, $B$, $C$ and $D$ are Hendrickson - Lattman coefficients. They constitute a complete record of the phase information for a single reflection. In case of several derivatives, Hendrickson - Lattman coefficients of all the reflections from all the derivatives are combined to get $A_{MIR}$, $B_{MIR}$, $C_{MIR}$ and $D_{MIR}$. Phase probability is calculated from combined $A_{MIR}$, $B_{MIR}$, $C_{MIR}$ and $D_{MIR}$ for every 15°.
Figure 4.4 Phase circles 112(a) and 317(b) reflections from Horse oxy-hemoglobin, illustrating the difference between most probable and the best phase. The curves show the corresponding phase distribution. The best phase $\alpha$ lies at the centroid of phase probability plot whereas; maxima denote the most probable phase $\alpha_1$. The structure factor corresponding to these phases are called as the best $F_p$ and most probable $F_{p}$. (Reprinted from Blundell and Johnson)
4.1.8 Statistics of Phasing

There are three major factors that give us the idea of the quality of phasing technique:

4.1.8.1 The Phasing Power

This is the ratio of the contribution of the heavy atom to the structure factor (|F_H|) and root mean square value of the lack of closure error (E),

\[ \frac{\langle |F_H| \rangle}{\langle E \rangle} \]

Phasing power of any derivative gives an estimate of the contribution from heavy atom to the phasing as well as the degree of derivatization. Ideally, a heavy atom compound with a phasing power >1 is thought to be a better contributor for phasing. It could be increased by improved structure solution, collecting better data sets or changing the soak conditions. It could also be used as a guide of maximum resolution at which the derivative can be used for phasing.

4.1.8.2 The Reliability Factors

This act as a guide for the heavy atom refinement. For acentric reflections, as discussed earlier the anomalous scattering factor is thought to be more accurate and the reliability factor is termed as R_kraut and it is calculated by

\[ \frac{\sum |F_{PH(obs)} - F_{PH(calc)}|}{\sum F_{PH(obs)}} \]

However, for centric reflections, the reliability factor is termed as R_cullis. It is a useful statistics for isomorphous phasing and it is calculated as given below,

\[ \frac{\sum ||F_{PH} \pm F_P| - F_{H(calc)}||}{\sum F_{PH} - F_{P}} \]

In the most ideal situation, the value of R_cullis should be between 0.5-0.7. Ideally, a MIR solution from a derivative with R_cullis value above than 0.7 is expected to be wrong and below 0.5 is an indication of an excellent derivative.
4.1.8.3 The Figure of Merit (FOM)

It is one of the most important parameters in phasing statistics. It is considered as a measure of precision of the best phase. An average FOM for entire map gives idea about the quality of the map. The FOM is the ratio between the $|F_{\text{P}}|_{\text{best}}$ and $|F_{\text{P}}|_{\text{obs}}$. In a phase probability circle, a more spread out curve will lead to poorer determination of $\alpha_{\text{P}}$ and thus the FOM will be low.

For MIR data, it is important to calculate a map weighted by FOM for each reflection in order to get cleaner and more interpretable map. Ideally, a map should have FOM value between 0.5 – 0.8 and maps with FOM < 0.5 are supposed to be noisy and difficult to interpret.

4.1.9 Molecular Replacement

One way to solve the phase problem is to compute phase using an atomic model obtained from the structure of a related protein. However, to build an atomic model in the new crystal form of the target protein, the known model should be correctly oriented and positioned in the new crystal lattice. Molecular replacement is the technique used to solve that problem. Molecular Replacement is the most effective and commonly used technique to solve the structure of a protein belonging to the same structural family. The level of resemblance of two protein structures correlates well with the level of sequence identity. As a rule of thumb, molecular replacement will probably be fairly straightforward if the model is complete and shares at least 40% sequence identity with the target sequence.

Like in MIRAS the Patterson function becomes useful in this method also. As discussed earlier, Patterson map is an interatomic vector map created by squaring the structure factor amplitudes and setting all phases to zero. This vector map contains a peak for each atom related to every other atom, with a large peak at origin indicating intra-atomic self vectors. A Patterson map is too noisy to derive any high resolution structural information. However, if Patterson maps are computed for the data derived from an unknown structure and from the structure of a previously solved homologue, in the correct orientation and position within the unit cell, the two Patterson maps should be closely correlated. It is possible by general operations involving rotation (orientation) and
translation (position), to bring the equivalent points of electron density in each molecule into coincidence.

If two position vectors $X_1$ and $X_2$ are equivalent points, then

$$X_2 = [C] X_1 + d$$

where $[C]$ is a rotation matrix and ‘$d$’ is a vector defining a translation. In this simplistic approach, the molecular replacement problem has three parameters to specify orientation and three to specify position and is typically divided into 2 steps: rotation and translation. Thus the procedure involved in MR is to find out the rotation and translation as a two-stage operation.

4.1.9.1 The Rotation Function

The Rotation function, exploits the fact that the intramolecular vectors depend only on the orientation of the molecule, not on its position in the unit cell. In the rotation function, the unknown Patterson map is compared to Patterson maps derived from the known homologue structure in different orientations. There will be one set for each molecular orientation found in the crystallographic cell. They must lie within the largest intermolecular distance ‘$r$’ from the origin of the Patterson function. The relative orientations of the molecules can be determined by cross-rotation function, which is computed by deriving the product function between the self-Patterson functions (model and target in crystal), which may be considered with in a sphere of radius ‘$r$’. At the relative orientation, the probe self-Patterson function will be superimposed on target in the crystal (Rossman & Blow 1962; Crowther, 1972). The product function $R$ is given by an equation,

$$R = \int P_2(X_2) P_1(X_1) \, dV$$

where $R$ is the product of Patterson $P_1$ and the rotated Patterson $P_2$ with in the volume $V$. This will have a higher value compared to all product fractions when the two self-Patterson functions are equivalently oriented and the highest correlation are obtained when the two structures (known and unknown) are in similar orientation(s). The $R$-factors and/or correlation coefficients are used to score the rotation function.

The magnitude of the rotation function is plotted in a three-dimensional space with the three angular rotations as the coordinates. Usually, Eulerian system is used as it
has the advantage over the other methods of showing the symmetry of the rotation function. The Eulerian angles could be defined by the angle \( \alpha \) around \( z \)-axis of an orthogonal coordinate system, then around the new \( y \)-axis by an angle \( \beta \), and finally a rotation by \( \gamma \) around the new \( z \)-axis (Figure 4.5). The sign of the rotation of the axes is positive for a clockwise rotation when looking from the origin along the positive rotation axis.

When two or more molecules exist in the asymmetric unit, a self-rotation function can be used to determine the orientation and an angle of rotation of non-crystallographic symmetry axis relating one molecule with the other.

![Eulerian system of angular rotation](image)

*Figure 4.5 Eulerian system of angular rotation.* It is used for determining spatial orientation of the known and unknown molecule with respect to each other in a rotation search.

### 4.1.9.2 The Translation Function

The Translation functions compare sets of intermolecular vectors, which depend on the position of the molecule in the unit cell with the observed Patterson. In the translation function, the correctly orientated known model can now be correctly positioned by translating it to the correct co-ordinates within the asymmetric unit. This is
accomplished by moving the model, calculating a new Patterson map, and comparing it to the unknown-derived Patterson map. Cross-Patterson vectors in this context are intermolecular vectors of the model structure in which two molecules are related by crystallographic symmetry. The computed cross-Patterson vectors \( P_{1,2}(u, t) \) obtained from the model structure are compared with observed Patterson function \( P(u) \). The translation function \( T(t) \) is calculated using the following equation (Rossman, et al., 1964; Crowther & Blow, 1967).

\[
T(t) = \int P_{1,2}(u, t) P(u) \, \text{du}
\]

A standard linear correlation coefficient 'C' is computed between observed structure factor amplitudes and the calculated structure factor from the translated molecules. The correlation coefficient is scaling insensitive and it is major advantage of this. The standard linear correlation coefficient \( C \) is

\[
C = \frac{\sum_{\text{all}} (|F_o|^2 - \text{Aver} |F_o|^2) (|F_c|^2 - \text{Aver} |F_c|^2) / \sum_{\text{all}} (|F_o|^2 - (\text{Aver} |F_o|^2)^2) \sum_{\text{all}} (|F_c|^2 - (\text{Aver} |F_c|^2)^2)^{1/2}}
\]

When the computed cross-Patterson fits correctly on the observed, it gives a higher value compared to all translation functions. Positions with high correlations are the outputs in cartesian coordinates.

4.1.9.3 Cell Content Analysis

Cell content analysis provides an estimate of the number of molecules in the asymmetric unit and thus the number of molecules to be searched in MR solution. Solvent generally occupies a part of the volume of protein crystals. Matthews's coefficient \( V_M \) is widely used to determine approximate volume of the contents of the asymmetric unit (Matthews, 1968). Matthews's coefficient indicates how much volume of the asymmetric unit contains a Dalton molecular weight of the protein and it is expressed in \( \text{Å}^3/\text{Da} \). It can be calculated as:

\[
\text{Volume of unit cell/ Molecular weight of protein} \times Z \times X
\]

where \( Z \) is the number of asymmetric units in the unit cell and \( X \) is the number of molecules in the asymmetric unit. The Matthews number is usually between 1.7 and 3.5 corresponding to a protein content of 75% to 30%. Information obtained from the cell
content analysis reduces the burden of searching for different number of molecules in the asymmetric unit.

4.1.10 Density Modification, Model Building and Refinement

After obtaining an initial map from MR or other experimental phasing techniques, it is modified by imposing restriction on the density in real space and this step is called as density modification. Iterative cycles of density modification improve the interpretability of the map. Density modification could be done by several ways like solvent flattening, histogram matching, NCS averaging etc. After density modification, experimental phases are improved by building the model and refining it by introducing some geometrical restraints and energy parameters.

Model building involves completing the model, fitting amino acid side chains in electron density map as well as adjustments of the bond length and bond angle. Refinement is the process of adjusting the model to find a closer agreement between the observed and the calculated structure factor. This process involves the refinement of three positional parameters (x, y, z) and an isotropic temperature factor (B) for all non-hydrogen atoms (in medium to high resolution structures) in order to improve the agreement between the observed |F₀| and the calculated |Fᶜ| structure factors. The occupancy is usually fixed as 1.0, unless the residue has alternate conformations. The hydrogen atoms refinement could be done when the structure is being solved at atomic resolutions.

The refinement methods in X-ray crystallography are based on the principle of least squares or maximum likelihood. In the least square refinement, observed structure factor amplitudes have fixed values and the parameters such as bond length, bond angle, planarity of aromatic rings and configuration of correct enantiomer are optimized to maximize the agreement between calculated and observed structure factor amplitudes. Gaussian distribution for probability distribution of structure factor amplitudes and phases are the principal requirement for least square refinement. Probability distributions of structure factors are Gaussian, however the phase information is lost when intensities are measured and therefore distribution of measurements no longer remains Gaussian. In addition, standard deviation of the Gaussian is an observed quantity and independent of
Chapter 4-Structure Determination, Model Building, Refinement and Validation of Pab-NTD and Pab-NTD-L-serine complex

the parameters of the model. Therefore, the conditions for least square refinement are not fulfilled; hence the least square results are not necessarily the best estimates of the model parameters.

4.1.10.1 Maximum Likelihood Method

The basic idea of maximum likelihood is that the best model is most consistent with the observations. (Read, 2001; Pannu & Read, 1996). The consistency is measured statistically, by the probability that the observations should have been made. The likelihood goes up if the model is changed to make the observations more probable, indicating that the model is better. The probabilities have to include the effects of all sources of error, including not just measurement errors but also errors in the model itself. This method overcomes the limitations observed in least square refinement. The principle of maximum likelihood method is based on Bayesian statistics, which is probability of the model given the data, \( p \{ \text{model}; \text{data} \} \), since data are the fixed observations.

\[
p[\{\text{model};\text{data}\}] = p[\text{model}] \frac{p[\text{data};\text{model}]}{p[\text{data}]}
\]

In crystallography, structure factor amplitudes \( |F_o| \) is the data and parameters are the parameters of the model. Calculated structure factor \( F_c \) is directly related to the model parameters; hence the above equation can be written as

\[
P[ F_c ; |F_o| ] = p [ F_c ] \frac{p[|F_o|; F_c]}{p [|F_o|]}
\]

4.1.11 Refinement Software

Crystallography and NMR System (CNS) version 1.0, software package (Brunger, et al., 1998) was used for macromolecular structure refinement in our investigations. CNS has several modules, which can be used for rigid body refinement, simulated annealing, minimization, B-factor and omit map calculation. Structure factor amplitude was used as refinement target in CNS. Maximum likelihood using structure factor amplitudes was used as the refinement method. Modules used in the refinement protocols are discussed below.
4.1.11.1 Rigid Body Refinement

Rigid body refinement is used as a constrained part rather than refining individual atomic positions as it considers the entire protein molecule or part of it as a rigid entity and its position in the unit cell is refined (Konnert, 1976). It could be used as a first step in the refinement procedure for accurate positioning of the molecule in the unit cell. This involves three rotation and three translation degrees of freedom of the rigid body.

4.1.11.2 Simulated Annealing

This method utilizes high temperature molecular dynamics and cooling to search stable peptide conformations. It involves randomizing and converting the solid state to liquid phase by heating protein molecule to a very high temperature (3500 K), followed by slow cooling of the system to 300 K at the rate of 25 K per cycle (Brunger, 1988; Brunger, et al., 1990). Simulated annealing is an energy minimization technique, particularly well suited to overcome the multiple minima problem. The method is based on VERLET minimizer algorithm, where the energy minimization is performed for “τ” seconds with the temperature of ‘T’ falling by δT in a time interval of δt (Verlet, 1967).

4.1.11.3 Conjugate Gradient Minimization

The refinement process may behave poorly on refining only the positions and B-factors of all the atoms. In that situation data will over fit and results in an inaccurate atomic model. This problem could be overcome either by adding "observations" in the form of restraints, or reduce the number of parameters by constraining the model in some way. Conjugate gradient minimization (E) is mainly used for stereochemically restrained least-squares refinement (Tronrud, et al., 1968). Conjugate gradient minimization consists of a crystallographic term (X-ray) and several stereochemical restraints.

\[ E_{\text{total}} = E_{\text{stereochemical}} + wE_{\text{xray}} \]

\( E_{\text{stereochemical}} \):

a) Restrains the distance between the atoms defining bond lengths, bond angles or dihedral angles.

b) Restrains the torsion angle, imposes certain restrictions on rotations around the Ramachandran Phi (\( \Phi \)) and Psi (\( \Psi \)) angles.
c) Imposes planarity on the aromatic ring.

d) Restrains the configuration around correct enantiomer.

\[ E_{\text{exray}} = \sum (|F_{o}(hkl)| - k|F_{c}(hkl)|)^2 \]

'w' is weighting term which can be adjusted according to the accuracy of the observation.

This is mainly used to relieve the short contacts as well as correct the geometry of the protein and it is based on Powell minimization (Powell, 1977). In the minimization process, restraints (\(E_{\text{stereochemical}}\)) are entered as the refinement target, and \(E_{\text{exray}}\) are weighted so that the deviations from ideal values match the deviations found in databases of high-resolution structures.

### 4.1.11.4 B-factor Refinement

Various static and thermal disorders can effectively "spread out" the electron density of a given atom, which in turn increases the B-factor. During structure refinement, B-factor could be refined in several ways like overall B-factor, the restrained and unrestrained B-factor, group B-factor and isotropic/anisotropic B-factor. In the present work, restrained individual isotropic B-factor of all atoms was refined. A significant missing part of a protein model consists of ordered water molecules hence after the initial model building and refinement of the protein part, water molecules were automatically picked at resolution >2.5 Å using the CNS task file ‘water_pick.inp’. These waters were considered as part of the protein structure and in the subsequent steps both the restrained individual isotropic B-factor of all atoms and coordinates for the water molecules were refined.

### 4.1.12 Fourier Maps and Map Interpretation

#### 4.1.12.1 2Fo-Fc and Fo-Fc Fourier Maps

After every cycle of refinement, the model was inspected and manually adjusted using double difference (2Fo-Fc) and difference (Fo-Fc) Fourier maps. 2Fo-Fc map is usually contoured at 1σ and used in model building whereas Fo-Fc map is contoured at...
Chapter 4-Structure Determination, Model Building, Refinement and Validation of Pab-NTD and Pab-NTD- L-serine complex

2.5-3.0σ and used in identification of missing parts of protein, solvent (ligand/inhibitor/water etc), mutated amino acids and so on. σ refers to the standard deviation from the mean electron density and is represented as the number of electrons/Å$^3$.

4.1.12.2 Omit maps

Simulated annealing omit maps are used to reduce the model biasness (Vijayan, 1980; Bhat & Cohen, 1984). The part of a model obtained by MR solution, which contains error could not be surely assigned, and therefore an omit map that covers the entire molecule is most useful. Composite omit map is calculated by systematically excluding small regions of the model. Phase informations from non-excluded part of the model are used for calculating map of the excluded part, thus reducing the model bias. These small maps are accumulated and written out as a continuous map covering the whole molecule (or the defined region). Simulated annealing refinement and minimization are used to remove the bias from the omitted region.

4.1.13 Validation

4.1.13.1 R-factor

The purpose of refinement is to make the model agree better with the observed diffraction data. The agreement of observed and calculated structure factors is often measured with the traditional R-factor (residual-factor or agreement factor), which is the average fractional disagreement:

\[ R = \sum (|Fo-Fc|)/\sum (Fo) \]

If we improve the agreement of observed and calculated structure factors assuming that the phases are correct, it is possible to overfit the data especially at moderate resolutions. Thus, refinement procedures generally tend to induce phase biasness, which is difficult to detect. This problem can be circumvented by using most of the data (working set) to refine the atomic model, and leaving the remaining data (test set) to verify the refinement protocol. The test set, containing a randomly selected set of reflections (~5%), is not used in the refinement of a structure and an R-free is computed, for this partial data set, in the same way as conventional R-factor (Brünger, 1992). In the
subsequent refinement steps, $R_{free}$ will be used as a cross-evaluator for the refinement and also to reduce the biasness in map calculations. Reduction in the value of $R_{free}$ with the advancement of refinement indicates that the model must have improved because the chance to overfit $R_{free}$ is very unlikely. Several cycles of model building and refinement are carried out till the $R$-factor reaches convergence or acceptable levels.

4.1.13.2 PROCHECK

Program 'PROCHECK' checks the stereo-chemical quality of protein structures (Laskowski, et al., 1993). It evaluates the structures on various geometric parameters and highlights the regions of the model that may need further investigation. These parameters mainly consist of torsion angles (main chain phi ($\Phi$)-psi ($\Psi$) dihedral angles and side chain Chi1-Chi2 torsion angles), peptide bond planarity, Cα-tetrahedron geometry distortion analysis, bond length and angle analysis and planarity of the aromatic residues. The output series consists of postscript files and comprehensive residue-by-residue analysis of geometric parameters. One example of 'PROCHECK' output is a Ramachandran plot. It is a way to visualize main-chain dihedral angles, phi ($\phi$) against psi ($\psi$), of amino acid residues in protein structure. It shows the possible conformations of $\phi$ and $\psi$ angles for a polypeptide and used to identify the residues in stereochemically allowed or disallowed region. For a well refined structure, most of the residues should lie in the allowed region of the Ramachandran plot. Additionally, 'PROCHECK' also compares and assess the quality of the model vis-à-vis other structures at comparable resolutions.

In this chapter the structure determination, model building, refinement and structure validation of the Pab-NTD and Pab-NTD-L-serine complex has been discussed.

4.2 Materials and Methods

4.2.1 Structure Determination of Pab-NTD

4.2.1.1 Scaling Derivative to Native Data Set

The 'IMPORT SCALE' task, as implemented in the CCP4 suite (CCP4. Acta Crystallogr., 1994) was used for the conversion of reflections from scalepack (sca) to CCP4 (mtz) format. The crystallographic reflections from various heavy atom derivative
and native data sets were merged and sorted by using program 'CAD' to generate a single data set. The merged datasets obtained by 'CAD' was scaled by SCALEIT and Fhscal program using Kraut's formula. Anomalous differences $F(+) \text{ and } F(-)$ for heavy atom data sets were included during scaling procedure. Five percent of randomly selected reflections are included in $R_{\text{free}}$ set.

4.2.1.2 Location of Heavy Atom Site and Phase Determination

The MIRAS structure was solved by automated phasing program 'SOLVE' (Terwilliger & Berendzen, 1999). It scales the data using local scaling, calculates difference Patterson maps and Fourier maps to find out heavy atom site, calculates the heavy atom model, refines the heavy atom parameters to get the best phase and finally looks for the presence of "solvent" and "protein" regions in the native electron density map. The 'SOLVE' output contains several files;

- solve.prt - Summary of SOLVE results
- solve.ezd - Portable electron density map
- solve.CCP4_map - CCP4 electron density map
- solve.mtz - CCP4 mtz file with F, Phases, Hendrickson-Lattman coefficients
- phases-hl.export - Formatted file with phases and Hendrickson- Lattman coefficients ready to read into CCP4
- phases-hl.script - Final parameters ready to use again in 'SOLVE'

The analysis of top solutions was scored in an output file 'solve.prt' which contains all the phasing informations and statistics. It also contains the coordinates of heavy atom sites in real space with its occupancies. These sites could also be located in the isomorphous difference Patterson map, computed using FFT, as implemented in CCP4 package. Various outlier filters, different resolution ranges and intensity differences were used for the calculation of Patterson maps to improve its interpretability. All the Patterson maps were visualized in the Harker section $z = 0.333$ and $0.667$. The heavy atom peaks found in the isomorphous difference Patterson could also be seen in the anomalous difference Patterson indicating that the structure solution was correct.
Chapter 4-Structure Determination, Model Building, Refinement and Validation of Pab-NTD and Pab-NTD- L-serine complex

4.2.1.3 The MIRAS Map and Density Modification

The output file solve.ezd contains portable electron density map for the MIRAS structure, which could be viewed in O version 9.0.2 (Jones, et al., 1991). This map requires further improvements to complete the structure solution. Before starting the refinement steps, the map is subjected for density modification procedure by using the program ‘RESOLVE’ (Terwilliger, 2000, 2003). It is an automated model-building program. In addition to density modification, it also performs a prime-and-switch phasing to remove model bias. By using the protein sequence, it looks for local pattern matching and identifies fragment to be fitted in the map to build a modified and better model.

4.2.1.4 Model Building and Refinement

The MIRAS map obtained from ‘SOLVE/RESOLVE’ was interpretable with distinct protein and solvent boundary. The phases were improved by using solvent flattening with the program ‘DM’ as implemented in CCP4. Initially, a poly-Alanine model was built for the structure, which in subsequent steps was changed to the respective amino acid residues. The MIRAS model of Pab-NTD was refined by 5 iterative cycles using CNS version 1.0 (Brünger, et al., 1998) by employing a maximum likelihood target function and simulated annealing based on torsion-angle dynamics. Each refinement cycle was followed by manual model rebuilding using mFo – dFc SIGMAA weighed maps by using ‘O’ version 9.0.2. In this process, the positions of previously unidentifiable residues could be built in the electron density map. The final refinement was done based on maximum likelihood positional refinement with restrained temperature factors by using two cycles of ‘REFMAC’ (Murshdov, et al., 1997). The model was adjusted in ‘O’ version 9.0.2 to improve geometrical parameters by using ‘REFIZONE’.

At each step of model building and refinement, the crystallographic R and R_free were monitored and refinement was carried out till convergence was achieved. The quality of the model was checked using ‘PROCHECK’ program that looks for the stereochemical quality of protein structures. CNS task file ‘water_pick.inp’ was used for adding water molecules to the structure and waters were picked at a level greater than
Chapter 4 - Structure Determination, Model Building, Refinement and Validation of Pab-NTD and Pab-NTD- L-serine complex

2.5σ from mF₀ – dF_c SIGMAA weighed maps. The water molecules thus picked were checked in ‘O’ version 9.0.2 before including them for further refinement.

4.2.2 Structure Determination of Pab-NTD-L-serine Complex

Pab-NTD-L-serine complex structure was solved using ‘MOLREP-AUTO MR’ as implemented in CCP4 suite (Vagin & Teplyakov, 1997) by using Pab-NTD structure as the probe. As discussed earlier, both phase as well as structure factor amplitude are required for determining the structure. In MR, the phase information is derived from the model and structure factor amplitude is derived from the ‘mtz’ file of a respective dataset. Like Pab-NTD, Pab-NTD-L-serine complex also belonged to P3_21 space group. Cell content analysis indicated the presence of one molecule per asymmetric unit and hence a search for one molecule in the asymmetric unit was carried out. The program was asked to search for 10 peaks in Rotation map and 10 in Translation function.

The density for the L-serine could be readily seen in difference Fourier maps. Further model building and refinement steps were carried out similarly as in the case of Pab-NTD. The MR model of Pab-NTD-L-serine complex was refined using CNS version 1.0 by employing simulated annealing based on torsion-angle dynamics and a maximum likelihood target function. Each refinement cycle was followed by model building using mF₀ – dF_c SIGMAA weighed maps by using O version 9.0.2. CNS task file ‘water_pick.inp’ was used for adding water molecules to the structure and the quality of the water molecules was checked in ‘O’ version 9.0.2. The validation of the structure was done by ‘PROCHECK’.
Chapter 4 - Structure Determination, Model Building, Refinement and Validation of Pab-NTD and Pab-NTD- L-serine complex

4.3 Results and Discussions

4.3.1 Structure Determination of Pab-NTD

4.3.1.1 Structure Solution by SOLVE

The structure solution of Pab-NTD was attempted by MIRAS method using the program ‘SOLVE/RESOLVE’ by using both P3121 as well as P3221 space groups for Pab-NTD. However, space group P3221 yielded the correct structure solution by ‘SOLVE’. The automated structure determination program yielded 7 sites for different heavy atom compounds. 1 site was occupied by gadolinium chloride, 3 sites for samarium acetate, 2 sites for thiomersal whereas sodium iodide occupied 1 site (Table 4.1). The peaks for all these sites could be located in the isomorphous difference Patterson as well as in the anomalous difference Patterson map except for the third site for samarium acetate and sodium iodide (Figure 4.6).

<table>
<thead>
<tr>
<th>SITE</th>
<th>ATOM</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>OCCUP</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gd</td>
<td>0.5543</td>
<td>0.2086</td>
<td>0.0000</td>
<td>0.9008</td>
<td>30.9075</td>
</tr>
<tr>
<td>2</td>
<td>Sm</td>
<td>0.2509</td>
<td>0.2543</td>
<td>0.0179</td>
<td>0.4787</td>
<td>36.8190</td>
</tr>
<tr>
<td>3</td>
<td>Sm</td>
<td>0.5384</td>
<td>0.2083</td>
<td>0.0052</td>
<td>0.6873</td>
<td>59.2603</td>
</tr>
<tr>
<td></td>
<td>Sm</td>
<td>0.1334</td>
<td>0.8849</td>
<td>0.0295</td>
<td>0.4225</td>
<td>60.0000</td>
</tr>
<tr>
<td>1</td>
<td>Hg</td>
<td>0.7424</td>
<td>0.7880</td>
<td>0.0603</td>
<td>0.4224</td>
<td>42.6259</td>
</tr>
<tr>
<td>2</td>
<td>Hg</td>
<td>0.9505</td>
<td>0.3168</td>
<td>0.1515</td>
<td>0.1172</td>
<td>42.0994</td>
</tr>
<tr>
<td>1</td>
<td>I</td>
<td>0.8816</td>
<td>0.8990</td>
<td>0.1137</td>
<td>0.0123</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Table 4.1 Positional coordinates of all heavy atom site given by ‘SOLVE’
Figure 4.6 Location of heavy atom sites in the Patterson maps. (a) Isomorphous difference Patterson map (3.5 Å) for gadolinium chloride (Gd) with the arrows indicating the Patterson peak (3σ) for heavy atom vector (0.10, 0.23). (b) Anomalous difference Patterson map for gadolinium chloride with the Patterson peak (5σ) marked at the same location.
Figure 4.6 Location of heavy atom sites in the Patterson maps. (c) Isomorphous difference Patterson map (3.5 Å) for samarium acetate (Sm) with the arrows indicating the three Patterson peaks (1σ and 2σ respectively) for heavy atom vector Sm1 (0.75, 0.25) and Sm2 (0.13, 0.25). (d) Anomalous difference Patterson map for samarium acetate with the corresponding Patterson peaks (both 3σ) marked at the same location.
Figure 4.6 Location of heavy atom sites in the Patterson maps. (e) Isomorphous difference Patterson map (3.0 Å) for thiomersal (Hg) with the arrows indicating the two Patterson peaks (4σ and 1σ respectively) for the heavy atom vector Hg1 (0.30, 0.47) and Hg2 (0.32, 0.22). (f) Anomalous difference Patterson map for thiomersal with the corresponding Patterson peaks (4σ and 2σ respectively) marked at the same location. Patterson peaks for the third site of samarium acetate and sodium iodide could not be located.
Phasing statistics indicated that sodium iodide could be a weak derivative as it has low phasing power as well as occupancy and relatively higher $R_{\text{culis}}$ (Table 4.2). However, MIRAS map obtained without incorporating sodium iodide data set turned out to be of poorer quality. ‘SOLVE’ could calculate a 2.5 Å MIRAS map with a mean figure of merit of 0.50 (Table 4.3).

<table>
<thead>
<tr>
<th>Heavy atom compounds</th>
<th>Phasing power</th>
<th>$R_{\text{culis}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Centric/Acentric</td>
<td></td>
</tr>
<tr>
<td>1. Gadolinium Chloride (Gd)</td>
<td>0.69/0.84</td>
<td>0.66</td>
</tr>
<tr>
<td>2. Samarium Acetate (Sm)</td>
<td>0.74/0.81</td>
<td>0.68</td>
</tr>
<tr>
<td>3. Thiomersal (Hg)</td>
<td>0.69/0.82</td>
<td>0.67</td>
</tr>
<tr>
<td>4. Sodium Iodide (I)</td>
<td>0.27/0.37</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Table 4.2 Phasing statistics for the various heavy atom compounds

<table>
<thead>
<tr>
<th>Figure of Merit</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of reflections</td>
<td>349</td>
<td>575</td>
<td>570</td>
<td>532</td>
<td>489</td>
<td>530</td>
<td>502</td>
<td>524</td>
<td>556</td>
<td>428</td>
</tr>
</tbody>
</table>

The MIRAS map has distinct solvent and protein regions (Figure 4.7). ‘SOLVE’ gives Z-scores for the Patterson cross-validation, difference Fouriers analysis of the native Fourier, figure of merit and overall scoring (Table 4.4). An overall Z-score should have a value of 5-10 for a 2-site solution or 20-50 for a 10-site solution. For this solution, the value of overall Z-score was 26.6, which indicated that structure solution could be correct. The summary of scoring for this solution is given below:

<table>
<thead>
<tr>
<th>Figure of Merit with Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{\text{min}}$</td>
</tr>
<tr>
<td>Number of reflections</td>
</tr>
<tr>
<td>mean Figure of Merit</td>
</tr>
</tbody>
</table>

Table 4.3 Figure of Merit (FOM) of the phasing solution given by ‘SOLVE’
Chapter 4-Structure Determination, Model Building, Refinement and Validation of Pab-NTD and Pab-NTD- L-serine complex

<table>
<thead>
<tr>
<th>Criteria</th>
<th>MEAN</th>
<th>SD</th>
<th>VALUE</th>
<th>Z-SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattersons</td>
<td>1.65</td>
<td>0.582</td>
<td>4.81</td>
<td>5.42</td>
</tr>
<tr>
<td>Cross-validation Fourier</td>
<td>4.33</td>
<td>3.97</td>
<td>73.7</td>
<td>17.5</td>
</tr>
<tr>
<td>NatFourier CCx100</td>
<td>10.7</td>
<td>4.78</td>
<td>34.3</td>
<td>4.94</td>
</tr>
<tr>
<td>Mean figure of meritx100</td>
<td>0.00E+00</td>
<td>9.89</td>
<td>50.3</td>
<td>5.09</td>
</tr>
</tbody>
</table>

|                                          | Correction for Z-scores | Overall Z-score values |
|                                          | -6.36                    | 26.6                   |

Table 4.4 Z-scores for the Patterson cross-validation, difference Fouriers analysis of the native Fourier, figure of merit and overall scoring given by ‘SOLVE’

4.3.1.2 Model Building and Refinement

The native Fourier map created by ‘SOLVE’ was subjected to density modification and model building by using the program ‘RESOLVE’. To calculate a modified map and to provide a model ‘RESOLVE’ requires the amplitude, experimental phase and Figure of Merit. All these informations are present in the ‘SOLVE’ output ‘solve.mtz’ which is utilized by ‘RESOLVE’. Additionally, it also requires solvent content of the crystal, protein sequence and heavy atom coordinates obtained by ‘SOLVE’. Once a modified map is calculated, it tries to trace a polypeptide chain and fit it in the map by using the libraries of actual helical/beta templates and side-chain densities to get a model. The solvent content for Pab-NTD crystal was 27.8% as calculated by using CCP4 suite of program. ‘RESOLVE’ calculated a 2.5Å map with a modified figure of merit (FOMM) of 0.68. In the automated model made by ‘RESOLVE’, 117 amino acid residues could be built in 7 fragments and side chains for 58 amino acids could be fixed (Figure 4.7b). The model covers a total of 52% of the asymmetric unit.
Figure 4.7 The MIRAS map and model for Pab-NTD. (a) The map had a clear solvent boundary and secondary structural elements could be readily observed. (b) The partial model of 117 residues as given by 'RESOLVE'. (c) The final refined Pab-NTD model with 143 residues.
Chapter 4-Structure Determination, Model Building, Refinement and Validation of Pab-NTD and Pab-NTD- L-serine complex

The 'RESOLVE' model was used for model completion and refinement. The MIRAS model of Pab-NTD was refined till 1.95Å resolution by initial 5 iterative cycles using CNS version 1.0 and finally by 2 cycles using ‘REFMAC’ (Table 4.5). During the fourth refinement cycle, water molecules were added to the structure. Using ‘O’ version 9.0.2, only genuine water molecules were selected in subsequent steps whereas spurious waters were removed. At each step of refinement, the values of $R_{\text{cryst}}$ and $R_{\text{free}}$ were monitored. The final model was refined to $R_{\text{cryst}}$ and $R_{\text{free}}$ values of about 20.6% and 27.2% respectively (Table 4.8). The Ramachandran plot computed to check the validity of the model indicated that 93.7% residues lies in the most favored region while the remaining 6.3% lie in additionally allowed region. None of the residues could be seen in the disallowed region.

<table>
<thead>
<tr>
<th>Resolution range (Å)</th>
<th>Overall R_factor (%)</th>
<th>Free R_factor (%)</th>
<th>R.M.S. Deviation</th>
<th>Average B Values (Å²)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Bond lengths (Å)</td>
<td>Bond angles (°)</td>
<td></td>
</tr>
<tr>
<td>Refinement using CNS version 1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 - 2.25</td>
<td>31.45</td>
<td>37.56</td>
<td>0.46</td>
<td>1.82</td>
</tr>
<tr>
<td>25 - 2.25</td>
<td>27.62</td>
<td>34.39</td>
<td>0.25</td>
<td>2.04</td>
</tr>
<tr>
<td>25 - 1.95</td>
<td>27.09</td>
<td>31.58</td>
<td>0.10</td>
<td>2.06</td>
</tr>
<tr>
<td>*25 - 1.95</td>
<td>23.78</td>
<td>32.01</td>
<td>0.01</td>
<td>2.17</td>
</tr>
<tr>
<td>25 - 1.95</td>
<td>22.79</td>
<td>32.29</td>
<td>0.01</td>
<td>2.22</td>
</tr>
<tr>
<td>Refinement using REFMAC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 - 1.95</td>
<td>20.65</td>
<td>27.44</td>
<td>0.01</td>
<td>1.34</td>
</tr>
<tr>
<td>25 - 1.95</td>
<td>20.55</td>
<td>27.22</td>
<td>0.01</td>
<td>1.32</td>
</tr>
</tbody>
</table>

Table 4.5 Refinement steps of Pab-NTD showing convergence of R-factor at different steps of refinement. Five cycles of refinement steps using CNS version 1.0 was followed by two cycles of refinement using ‘REFMAC’. Each refinement cycle was followed by manual adjustment of model using ‘O’ version 9.0.2. * Water molecules were added to the model after the third step of refinement.
Figure 4.8 The Ramachandran plot for Pab-NTD structure obtained using ‘PROCHECK’
4.3.2 Structure Determination of Pab-NTD-L-serine Complex

4.3.2.1 Structure Solution by MR

Pab-NTD-L-serine complex structure was solved using ‘MOLREP-AUTO MR’ version 7.5.01, as implemented in CCP4 suite, by using Pab-NTD structure. The ‘Cell Content Analysis’ module from the CCP4 suite gave a Matthews coefficient ($V_M$) of 2.8 Å$^3$/Da for the crystals and the solvent content of the crystal was predicted to be 27.8%. This indicated that the crystal contained one molecule in the asymmetric unit.

The Rotation solutions are defined using both Euler and Polar angles. The Euler angles, Polar angles, RF and RF/$\sigma$ of the highest solution with other solutions are shown in Table 4.6a. The statistical parameters used to select the correct Rotation and Translation solution should have significantly higher values when compared to the rest of the solutions. As can be noted, RF/$\sigma$ of the best solution is significantly higher than the rest of the solutions.

<table>
<thead>
<tr>
<th>Polar Angles</th>
<th>Euler angles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>theta Phi Chi</td>
</tr>
<tr>
<td>Sol_RF 1</td>
<td>179.03 -139.59 178.84</td>
</tr>
<tr>
<td>Sol_RF 2</td>
<td>69.95 1.62 178.89</td>
</tr>
<tr>
<td>Sol_RF 3</td>
<td>95.57 178.77 113.96</td>
</tr>
</tbody>
</table>

Table 4.6a The Rotation solution of Pab-NTD-L-serine complex as given by MOLREP. The Polar angles, Euler angles, RF and RF/$\sigma$ of the highest solution with other solutions are shown. The program was asked to search for 10 peaks for Rotation function (Only top 3 Rotation solutions have been shown).

All the 10 Rotation solutions are used for computing the Translation solutions. TF/$\sigma$, R-factor and correlation coefficient parameters are computed in a translation search. The fractional coordinates in all three axes are used to indicate the Translation solution. The program uses correlation coefficient as the criterion for the selection of the best Translation solution, since it is insensitive to scaling between observed and calculated structure factor amplitudes. Invariably, for the correct solution all these three
parameters are expected to be better than other solutions (Table 4.6b). The contrast observed between the highest Rotation peak solution and rest of the solutions indicated that a clear Rotation and Translation solution is obtained as shown in Table 4.6c.

Table 4.6b The Rotation and Translational functions of Pab-NTD-L-serine complex as given by MOLREP. The program was asked to search for 10 peaks in Rotation function and 10 in Translation function. Solutions of top 3 Translation functions for top 3 Rotation functions are given. The statistical parameters of the correct Rotation and Translation solution have significantly higher values when compared to the rest of the solutions.
Chapter 4-Structure Determination, Model Building, Refinement and Validation of Pab-NTD and Pab-NTD- L-serine complex

<table>
<thead>
<tr>
<th>Total number of monomers fixed: 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha</td>
</tr>
<tr>
<td>Sol_Mon_1_Rf_1_Tf_1</td>
</tr>
</tbody>
</table>

Table 4.6c The highest Rotation and Translational function of Pab-NTD-L-serine complex as given by MOLREP. The contrast observed between the highest Rotation and Translation solution and the rest of the solution indicating the correctness of solution.

4.3.2.2 Model Building and Refinement

The Pab-NTD-L-serine complex structure was refined using CNS version 1.0, as discussed in Section 4.2.2. Each refinement cycle was followed by manual model building using mF_o – dF_c SIGMA_A weighed maps using ‘O’ version 9.0.2. The model obtained from MR was subjected for rigid body refinement followed by energy minimization using the ‘CNS’ task file ‘anneal.inp’ and atomic B-factor refinement. Water molecules were added to the structure by task ‘water_pick.inp’. Fo-Fc and 2Fo-Fc Fourier maps were calculated from ‘CNS’ task file ‘model_map.inp’. These maps were used for analyzing the quality of the model (it includes protein, water as well as ligand, if any) and manual inspection of electron density was performed using ‘O’ version 9.0.2.

A clear density for L-serine could be observed in difference Fourier maps. First cycle of refinement using CNS was followed by model building in ‘O’ version 9.0.2. The models built from ‘O’ version 9.0.2 were subjected to few cycles of conjugate gradient restrained minimization, individual B-factor refinement using the ‘CNS’ task files ‘minimization.inp’ and ‘bindividual.inp’, respectively. At the third step of refinement L-serine was added to the structure and once again it was subjected to energy minimization by using the ‘CNS’ task file “anneal.inp” followed by atomic B-factor refinement. Finally, two cycles of CNS refinement was carried out by using conjugate gradient minimization, followed by individual B-factor refinement. L-serine could be nicely fitted in the difference Fourier maps (Figure 4.9).
Figure 4.9 L-serine bound in the active site of Pab-NTD. A clear density for L-serine could be seen in the difference Fourier maps. A (2Fo-Fc) map, contoured at 1σ, shown around the bound L-serine after the final refinement.
Chapter 4-Structure Determination, Model Building, Refinement and Validation of Pab-NTD and Pab-NTD-L-serine complex

At each step of refinement, the values of $R_{\text{cryst}}$ and $R_{\text{free}}$ were monitored and refinement was carried out until convergence was reached (Table 4.7). After the final cycle, $R_{\text{cryst}}$ and $R_{\text{free}}$ reached respectable values of 21.96% and 28.59% respectively (Table 4.8). In the final model of Pab-NTD-L-serine complex structure, 90.6% of the residues are present in the most favoured region and none in disallowed region of the Ramachandran Plot (Figure 4.9).

<table>
<thead>
<tr>
<th>Stages</th>
<th>Refinement protocol</th>
<th>$R_{\text{cryst}}$ (%)</th>
<th>$R_{\text{free}}$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Rigid body refinement followed by Energy minimization (annealing) and atomic B-factor refinement</td>
<td>26.59</td>
<td>31.42</td>
</tr>
<tr>
<td>2.</td>
<td>Water picking followed by restrained minimization</td>
<td>23.82</td>
<td>32.14</td>
</tr>
<tr>
<td>3.</td>
<td>Atomic B-factor refinement</td>
<td>23.07</td>
<td>30.97</td>
</tr>
<tr>
<td>4.</td>
<td>Energy minimization (annealing) with water molecules and ligand (L-serine)</td>
<td>23.03</td>
<td>30.47</td>
</tr>
<tr>
<td>5.</td>
<td>Atomic B-factor refinement</td>
<td>22.74</td>
<td>30.05</td>
</tr>
<tr>
<td>6.</td>
<td>Restrained minimization followed by atomic B-factor refinement</td>
<td>21.99</td>
<td>29.01</td>
</tr>
<tr>
<td>7.</td>
<td>Restrained minimization followed by final atomic B-factor refinement</td>
<td>21.96</td>
<td>28.59</td>
</tr>
</tbody>
</table>

Table 4.7 The Refinement steps of Pab-NTD-L-serine complex showing convergence of R-factor after different steps of refinement. Each refinement cycle was followed by manual adjustment of model using ‘O’ version 9.0.2.
Figure 4.10 The Ramachandran plot for Pab-NTD-L-serine complex structure obtained using ‘PROCHECK’
Chapter 4 - Structure Determination, Model Building, Refinement and Validation of Pab-NTD and Pab-NTD- L-serine complex

<table>
<thead>
<tr>
<th>Refinement Statistics</th>
<th>Pab-NTD</th>
<th>Pab-NTD-L-serine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution (Å)</td>
<td>1.95</td>
<td>2.10</td>
</tr>
<tr>
<td>No. reflections</td>
<td>9957</td>
<td>8399</td>
</tr>
<tr>
<td>R (%)</td>
<td>20.60</td>
<td>21.96</td>
</tr>
<tr>
<td>R&lt;sub&gt;free&lt;/sub&gt; (%)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>27.20</td>
<td>28.59</td>
</tr>
<tr>
<td>No. residues</td>
<td>143</td>
<td>143</td>
</tr>
<tr>
<td>Number of atoms</td>
<td>1316</td>
<td>1253</td>
</tr>
<tr>
<td>Protein</td>
<td>1140</td>
<td>1140</td>
</tr>
<tr>
<td>L-serine</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Water</td>
<td>176</td>
<td>106</td>
</tr>
<tr>
<td>R.m.s. deviation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bond length (Å)</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td>Bond angle (°)</td>
<td>1.32</td>
<td>1.41</td>
</tr>
<tr>
<td>Mean B value (Å²)</td>
<td>33.3</td>
<td>29.5</td>
</tr>
</tbody>
</table>

Table 4.8 The summary of refinement statistics for Pab-NTD as well as Pab-NTD-L-serine complex. *Throughout the refinement, 5% of the total reflections were held aside for R<sub>free</sub>.