DIRECT PHAING OF PROTEIN CRYSTALS WITH HIGH SOLVENT CONTENT

A Dissertation Presented to

the Faculty of the Department of Physics

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Hongxing He

May 2015

DIRECT PHAING OF PROTEIN CRYSTALS WITH HIGH SOLVENT CONTENT

Hongxing He

APPROVED:

Dr. Wu-Pei Su , Chairman

Dr. Gemunu Gunaratne

Dr. Pei-Herng Hor

Dr. David R. Jackson

Dr. Chin-Sen Ting

Dean, College of Natural Sciences and Mathematics

Acknowledgements

First of all, I would like to express my sincere gratitude to Dr. Wu-Pei Su, my dissertation advisor, for his constant support, patience, encouragement, and freedom he gave me through the past years. His technical and editorial advice was essential to the completion of this dissertation. I have learned a lot from him during these years. His insight in this area always enlightens me.

My thanks also go to other members of my dissertation committee, Dr. Gemunu Gunaratne, Dr. Pei-Herng Hor, Dr. David R. Jackson, and Dr. Chin-Sen Ting for providing many valuable comments on the presentations of Annual Progress Evaluation.

I want to thank my colleagues: Hoyin Chan, Hengrui Fang, Bo Li, Wei Li, Guoxiong Su, Yuan-Yen Tai, and Yuanyuan Zhao. I thank Dr. Yu-Hui Dong for useful correspondence and Dr. George Phillips Jr for useful conversations.

I would also like to thank Dr. Kevin E. Bassler, he helped me a lot when I came to UH. I want to thank Dr. George Reiter. He always encourages students to ask questions in his class. I also want to thank Chin-Davis Jennifer and Naomi Haynes.

Last, but not least, my deepest love goes to my parents, my brother, and my sister. Their love and support gave me strength throughout the completion of my studies and this dissertation.

This work was partially supported by the Texas Center for Superconductivity and the Robert A. Welch Foundation (E-1070).

DIRECT PHAING OF PROTEIN CRYSTALS WITH HIGH SOLVENT CONTENT

An Abstract of a Dissertation

Presented to

the Faculty of the Department of Physics

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

By

Hongxing He

May 2015

Abstract

Determining the phases of a diffraction pattern is crucial since the diffraction pattern of a protein crystal yields only the magnitude of the Fourier transform of the electron density. In order to invert the diffraction pattern to get the protein structure, the phase problem must be solved.

An iterative transform method is proposed for solving the phase problem in protein crystallography. In each iteration, a weighted average electron-density map is constructed to define an estimated protein mask. Density modifications are then imposed through the histogram matching technique in the protein region, and the hybrid input–output algorithm in the solvent region. Starting from random initial phases, after thousands of iterations the calculated protein mask evolves into the correct shape and the phases converge to the correct values with an average error of $30^\circ \sim 40^\circ$ for high-resolution data for several protein crystals with high solvent content. With the use of non-crystallographic symmetry and other density constraints, the method could potentially be extended to phase protein crystals with less than 50% solvent fraction. The new phasing algorithm can supplement and enhance the traditional refinement tools.

Contents

Acknowledgements iii

Abstract vi

Contents vii

1 Introduction to protein X-ray crystallography 1

1.1 Introduction 1

1.2 Gene cloning, DNA expression, and protein purification 3

1.3 Protein crystallization 5

1.4 X-ray diffraction, data collection, and indexing 8

1.5 Protein model building and model refinement 15

1.6 Summary 20

2 Principles of protein X-ray crystallography 22

2.1 Introduction 22

2.2 X-ray scattering by electrons 23

2.3 X-ray diffraction by a crystal . . . 26
2.4 Fourier transform of the diffraction pattern . . . 28
2.5 Temperature factor . . . 30
2.6 Atomic radius . . . 31
2.7 Synthetic diffraction data of a protein crystal . . . 32
2.8 Summary . . . 37

3 Methods for solving the phase problem **39**

3.1 Introduction . . . 39
3.2 Patterson function . . . 40
3.3 Isomorphous replacement method . . . 42
3.4 Anomalous dispersion method . . . 46
3.5 Molecular replacement method . . . 52
3.6 Direct method . . . 55
3.7 Summary . . . 61

4 Iterative transform method **62**

4.1 Introduction . . . 62
4.2 Oversampling condition . . . 64
4.3 Flowchart of the iterative transform method . . . 65
4.4 Symmetry operations, equivalent positions, and origin choices . . . 69
4.5 Fast Fourier Transform . . . 76
4.6 Weighted average density . . . 80
4.7 Histogram matching . . . 86

4.8 Hybrid input-output . . . 89
4.9 Solvent flattening . . . 91
4.10 Missing reflections . . . 91
4.11 R factor . . . 93
4.12 Summary . . . 94

5 Direct phasing of protein crystals with high solvent content 96

5.1 Introduction . . . 96
5.2 A photosynthetic reaction center structure with PDB ID 2UXJ . . . 98
5.3 A formaldehyde-activating enzyme (Fae) structure with PDB ID 1Y5Y . . 105
5.4 A human thyroid hormone receptor with PDB ID 3ILZ . . . 113
5.5 A pig pancreatic alpha-amylase with PDB ID 1WO2 . . . 119
5.6 A flavor protein WrbA from *Escherichia coli* with PDB ID 3B6I . . . 124
5.7 Discussions . . . 134
5.8 Conclusions . . . 135

6 Direct phasing of protein crystals with low solvent content 137

6.1 Introduction . . . 137
6.2 NCS averaging . . . 139
6.3 Gradient-histogram matching . . . 145
6.4 Conclusions . . . 155

Bibliography 156

Chapter 1

Introduction to protein X-ray crystallography

1.1 Introduction

X-ray crystallography can be essentially treated as a form of very high resolution microscopy . It enables us to view the protein structures at the atomic level and enhances our understanding of the protein functions. For example, we can study the interaction between proteins and other molecules, the conformational changes of proteins, and the catalysis mechanisms in the case of enzymes.

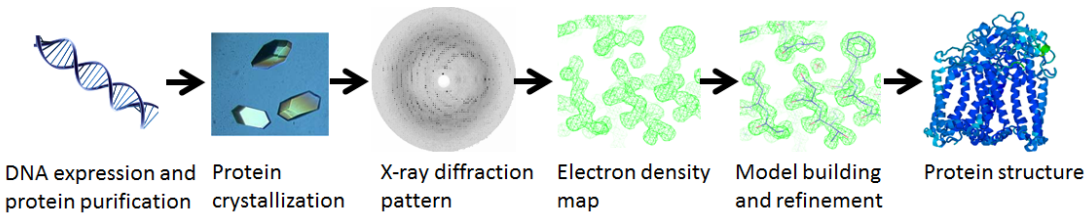

FIGURE 1.1: Schematic diagram of protein X-ray crystallography.

Workflow for solving the structure of a protien by X-ray crystallography is shown in Figure 1.1[1–3]. Protein can be expressed by messenger RNA which translates genetic information from DNA and protein has to be purified before crystallization. Protein X-ray crystallography includes protein crystallization, X-ray diffraction, solving the phase problem, model building, model refinement, and interpretation. Protein crystallization is the most challenging part in protein X-ray crystallography. After crystallization, a protein crystal is mounted onto a goniometer head for X-ray diffraction experiment. When the crystal on the goniometer head rotates, series of diffraction patterns are recorded and indexed until the entire diffraction data have been collected. On a diffraction pattern only the magnitudes of the structure factors are recorded, the phases of the structure factors are lost. This is called the phase problem, which constitutes another challenge in protein X-ray crystallography. After solving the phase problem, model structure could be built into the calculated electron density map. Refinement makes the model structure more accurate.

This chapter is organized as follows. In Section 1.2, a short introduction is given about gene cloning, DNA expression, and protein purification. In Section 1.3, protein crystallization is described. In Section 1.4, X-ray diffraction, data collection, and indexing are introduced. In Section 1.5, protein model building and model refinement are discussed. The last section is a short summary of protein X-ray crystallography. X-ray diffraction theory will be introduced in Chapter 2 and the phasing methods will be discussed in Chapter 3.

1.2 Gene cloning, DNA expression, and protein purification

Protein expression is important in drug discovery. For example, proteins can be screened as biological targets or as potential drugs. Protein expression also has significant applications in industry, such as the manufacture of enzymes and the production of human insulin to treat diabetes. In order to produce a large amount of proteins, gene cloning is required.

Gene cloning can provide unlimited quantity of a gene of interest[4]. Gene cloning needs a gene of interest and a vector which will carry the gene of interest. DNA containing the gene of interest is taken from its cell. Small circular DNA molecules called plasmids are removed from bacterial cells. These plasmids serve as vectors and they will carry the gene of interest.

A restriction enzyme can recognize the specific restriction sites on the DNA sequence. It can locate the gene of interest from its DNA. It can also open the circular plasmids. The gene of interest gets included into some of the opened plasmids, forming the recombinant plasmids. DNA ligase makes the combination permanent. The recombinant plasmids are mixed with the becteria. Some of them take up the plasmids in a process called transformation. Those bacteria with recombinant plasmids can be identified and be allowed to reproduce. The gene of interest on the recombinant plasmid is cloned.

Proteins are expressed from cloned DNA, as shown in Figure 1.2. First, the cloned DNA is transcribed to a messenger RNA. Then the messenger RNA is translated into

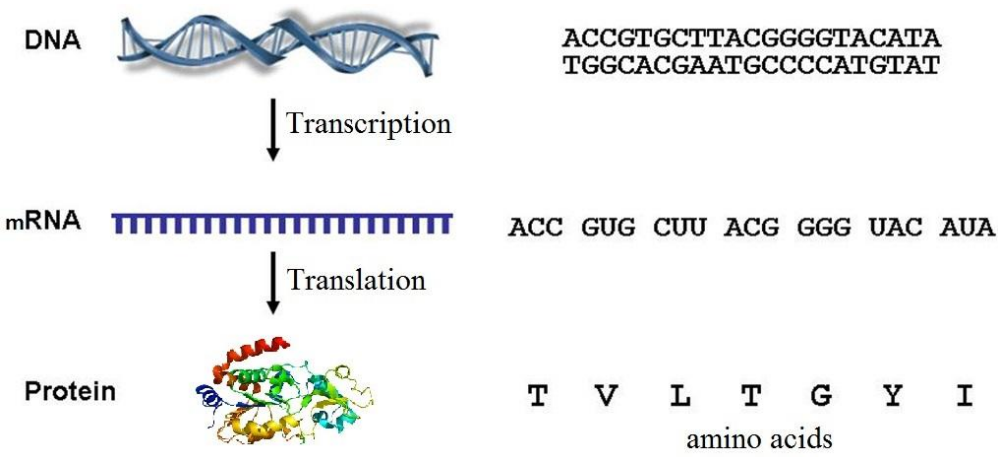

FIGURE 1.2: The central dogma of gene expression from DNA to mRNA to protein.

polypeptide chains. The polypeptide chains are ultimately folded into protein molecules. The protein peptide sequences are available in FASTA format in the Protein Data Bank, in which amino acids are represented by single-letter codes.

There are many ways to get the cloned DNA expressed in a host cell. Various host cells can be used for DNA expression. Expression systems are referred to by the host and the cloned DNA sources. Common hosts include bacteria, yeast, and eukaryotic cells. Common DNA sources include plasmids, viruses, bacteriophage, and artificial chromosomes. Since large amounts of protein molecules are needed for protein crystallization, bacterial expression is often used.

Protein purification is vital for the characterization of protein structure and function. The purification methods can be roughly divided into analytical and preparative methods. Preparative methods aim to produce large quantities of proteins which are commonly used in structural biology. The presence of impurities can affect the protein crystal growth. Generally, the purer the protein is, the easier to grow crystals.

1.3 Protein crystallization

Protein crystallization is the process of growing a protein crystal[1–3]. Protein molecules can form crystals when the solution in which they are dissolved gets supersaturated. Individual molecules can pack into a periodic array by noncovalent interactions.

Before crystallization, protein is dissolved in solvent. The solvent should be suitable for the protein to be dissolved and precipitated in crystalline form. The solvent is usually a water-buffer solution containing little or no salt. Sometimes, the solvent is a water-organic solution, with 2-methyl-2,4-pentanediol (MPD) added. Membrane proteins require water-detergent solution.

The precipitant solution is then added to a concentration that a precipitate does not develop. The precipitant solution is usually water-salt solution or water-polyethyleneglycol (PEG) solution. The most popular salt is ammonium sulfate which has a high solubility in water.

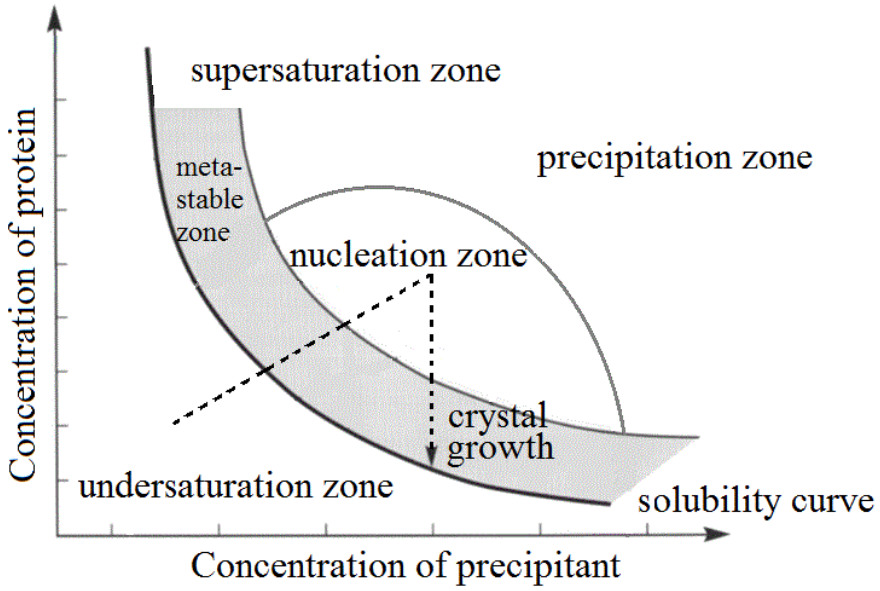

FIGURE 1.3: Phase diagram for protein crystallization mediated by a precipitant, and the ideal strategy (the dash line) for growing big crystals.

The ideal strategy for growing protein crystals is shown in Figure 1.3. Slowly increase the concentration of the precipitant such as salt, PEG, or organic solvent to make the protein solution reach supersaturation. When supersaturation reaches high level, spontaneous formation of protein nuclei is best achieved. High supersaturation may create too many small nuclei and therefore too many tiny crystals. In order to get big crystals, the crystals should grow slowly to obtain a best degree of order. After nuclei are formed, the supersaturation should be reduced to a lower level to make the nuclei slowly grow to big crystals. In practice, by changing the pH or the temperature, the supersaturation level of the protein solution can be changed.

A great number of trial experiments should be carried out at the same time in order to find the best crystallization conditions. A reasonable size of a protein crystal is about $0.3mm \times 0.3mm \times 0.3mm$.

Protein crystals include protein and some solvent. The average electron density in the protein region is about $0.43e/A^3$ while the average electron density in the solvent region ranges from $0.33e/A^3$ for pure water to $0.41e/A^3$ for salt, such as 4M ammonium sulphate.

In practice, several crystallization methods are used, such as vapor diffusion and microdialysis. Protein solution and precipitant solution are separated. Water molecules come out from the protein solution and go to the precipitant solution. Hence the protein solution becomes supersaturated and nucleation occurs.

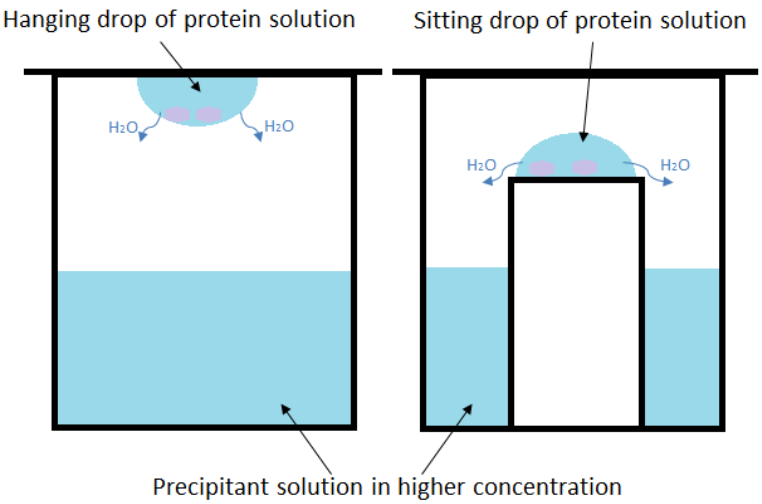

FIGURE 1.4: Growing protein crystals by the hanging-drop and the sitting-drop methods.

Vapor diffusion is often used to grow protein crystals. It includes the hanging-drop and the sitting-drop methods. The purified protein is dissolved into a buffer containing precipitant, such as salt or polyethyleneglycol (PEG). A drop of protein and precipitant solution is suspended to the downside of a cover slip or sits on the top of a small island inside a container, as shown in Figure 1.4. The container is filled with a reservoir of precipitant solution which has an optimal precipitant concentration for producing crystals. The droplet hangs on top of the precipitant reservoir or sits on the island in the reservoir. The container is sealed. Water vapor comes out from the droplet which becomes supersaturated and protein nucleation occurs. Because the container is sealed, as nucleation occurs, the protein concentration in the droplet decreases. The nucleation stops and the existing nuclei grow into small crystals which can be used as seeds to grow big crystals that are large enough for diffraction experiment.

Microdialysis is another method used in protein crystal growth. A membrane is used to separate the protein solution and the precipitant solution, as shown in Figure 1.5.

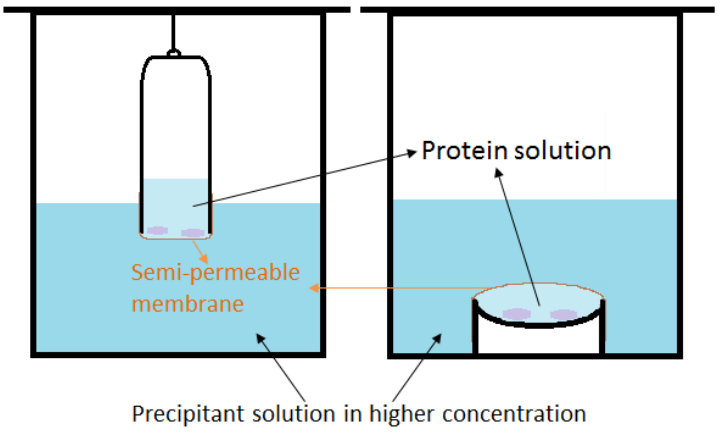

FIGURE 1.5: Growing protein crystals by the microdialysis method.

The membrane is semi-permeable. Small molecules such as water molecules can pass the membrane while protein molecules can't cross the membrane. Because of the higher concentration of the precipitant solution, water molecules come out from the protein solution and go to the precipitant solution. The protein solution becomes supersaturated and protein nuclei are prompted to grow.

1.4 X-ray diffraction, data collection, and indexing

The diffraction from a single molecule is too weak to be measurable. Protein crystal has a repeating formation of protein molecules. The observed diffraction pattern is a superposition of many diffractions from identical protein molecules. The observed intensities become strong when the crystal is big. In order to introduce the X-ray diffraction pattern, real space and reciprocal space lattices should be introduced first.

A crystal can be described in terms of its unit cell. The crystal lattice can be thought of as an array of unit cells by translation. A diffracted beam can be treated as a reflection from a set of equivalent, parallel planes of atoms. In particular, a family

of lattice planes is denoted by three integers h, k, and l, the Miller indices. They are written as (hkl) and represent the family of parallel planes orthogonal to $h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$, where $\mathbf{a}^*$, $\mathbf{b}^*$, and $\mathbf{c}^*$ are the basis of the reciprocal lattice vectors. The reciprocal lattice corresponding to the crystal lattice can be defined in the reciprocal space.

$$\begin{aligned} \mathbf{a}^* &= 2\pi \frac{\mathbf{b} \times \mathbf{c}}{\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})} \\ \mathbf{b}^* &= 2\pi \frac{\mathbf{c} \times \mathbf{a}}{\mathbf{b} \cdot (\mathbf{c} \times \mathbf{a})} \\ \mathbf{c}^* &= 2\pi \frac{\mathbf{a} \times \mathbf{b}}{\mathbf{c} \cdot (\mathbf{a} \times \mathbf{b})} \end{aligned} \tag{1.1}$$

$\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are the three lattice vectors that define the crystal unit cell.

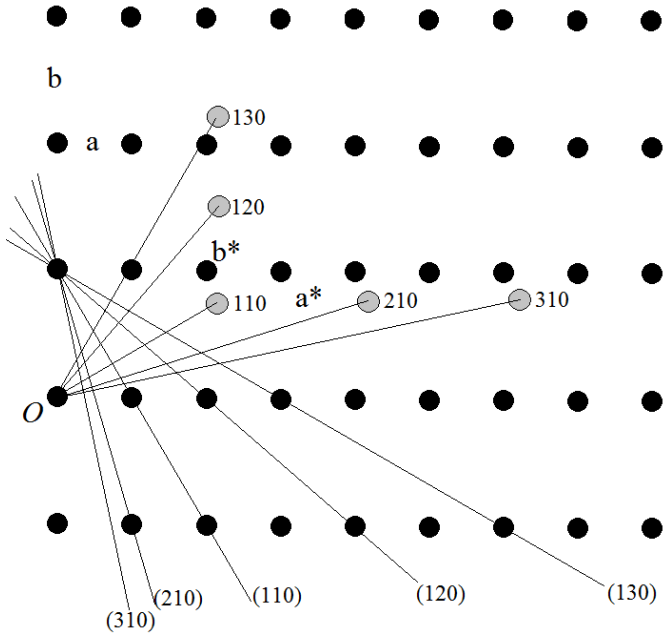

FIGURE 1.6: Construction of reciprocal lattice (gray) from real lattice (black).

According to its definition, a reciprocal lattice can be constructed from the real-space crystal lattice, which is shown in Figure 1.6. O is the origin of the reciprocal

lattice. Reciprocal space vectors $h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$ are perpendicular to the parallel planes denoted by the Miller indices h, k, and l.

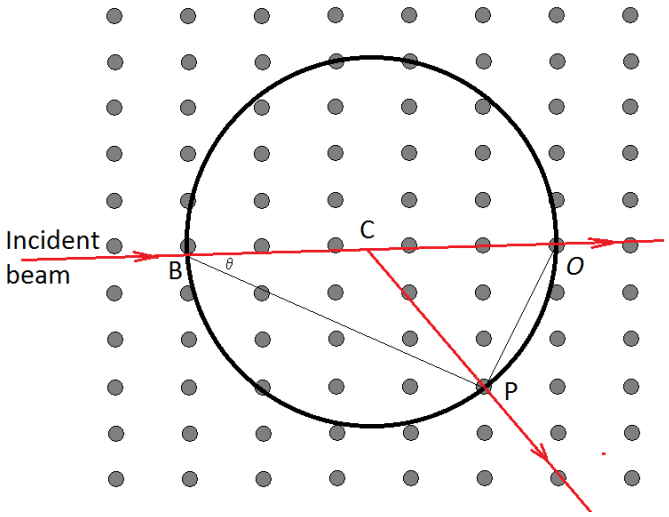

FIGURE 1.7: Ewald's sphere and a diffraction in reciprocal space.

In reciprocal space, a diffraction can be described by the Ewald sphere with a radius $1/\lambda$. In Figure 1.7, O is the origin of the reciprocal lattice. P is a point on reciprocal lattice and the length of OP is $1/d_{hkl}$.

$$\sin\theta = \frac{OP}{OB} = \frac{1/d_{hkl}}{2 \times 1/\lambda} \tag{1.2}$$

$$2d_{hkl}\sin\theta = \lambda \tag{1.3}$$

Equation 1.3 is Bragg's law. Ewald's sphere can be used to find the maximum resolution available for a given X-ray wavelength and the unit cell dimensions.

The crystal is mounted on the head of a goniometer. When the crystal rotates in the X-ray beam, various reciprocal-lattice points come into contact with the surface of

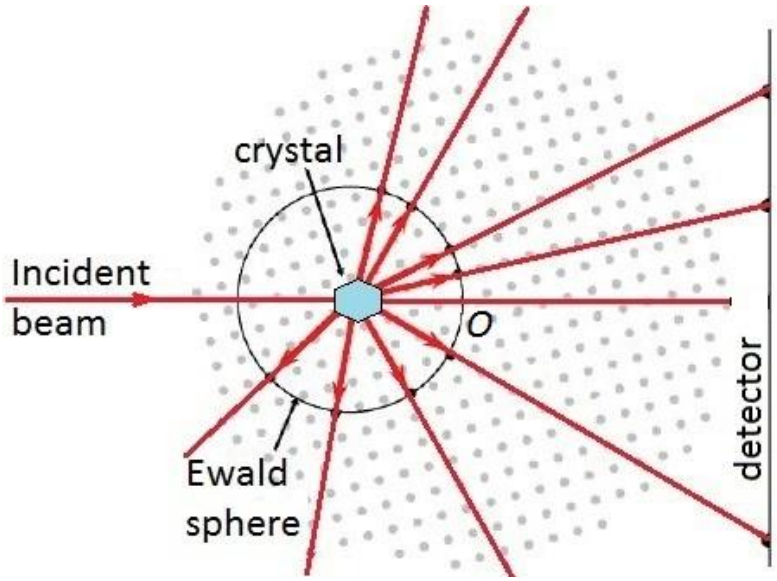

FIGURE 1.8: When a reciprocal-lattice point intersects the Ewald sphere, a diffraction ray emerges from the protein crystal as a reflection.

the Ewald sphere, and lots of diffracteded rays are produced in the direction of lines from the center of the Ewald sphere through the reciprocal-lattice points, as shown in Figure 1.8. These rays are recorded on the detector as a diffraction pattern. When the reciprocal-lattice point P_{hkl} is in contact with the Ewald sphere, the diffraction spot produced is called the hkl reflection, because it can be represented as a reflection by a set of equivalent, parallel, real-space planes with Miller indices (h, k, l) according to Bragg's law. The directions of diffracted rays only depend on the dimensions of the real-space unit cell and the wavelength of the incident beam. The electron density distribution inside the unit cell determines the diffraction intensities.

Figure 1.9 shows a picture of X-ray diffraction system used in the lab. It consists of X-ray source, goniometer head, diffraction pattern detector, beam stop, and N_2 stream channel. The X-ray source is a conventional X-ray tube and it provides the incident beam. The crystal is mounted on the goniometer head and cooled by the N_2 stream. The

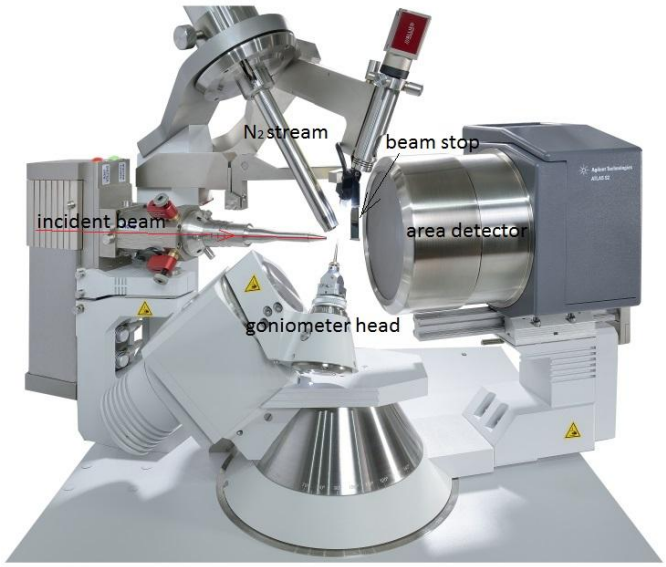

FIGURE 1.9: X-ray diffraction instrument used in the lab.

goniometer head can rotate in three angles. When the crystal rotates, the corresponding reciprocal lattice also rotates. Various reciprocal-lattice points get contact with the surface of the Ewald sphere and diffracted rays are recorded on a diffraction pattern by the area detector. Each rotation of the crystal gives a diffraction pattern. The entire set of diffraction data is the combination of all diffraction patterns.

A beam stop is always needed to stop the intense incident beam that has not been diffracted by the crystal. Otherwise, the detector might be damaged. Usually the beam stop can be completely impenetrable to the X-ray. The beam stop results in the absence of some diffraction spots with small diffraction angles on the diffraction pattern.

In addition to the conventional X-ray tubes, synchrotron light sources are also used in X-ray crystallography. The synchrotron light source is a source of electromagnetic radiation. It is usually produced by a storage ring and specialized particle accelerators

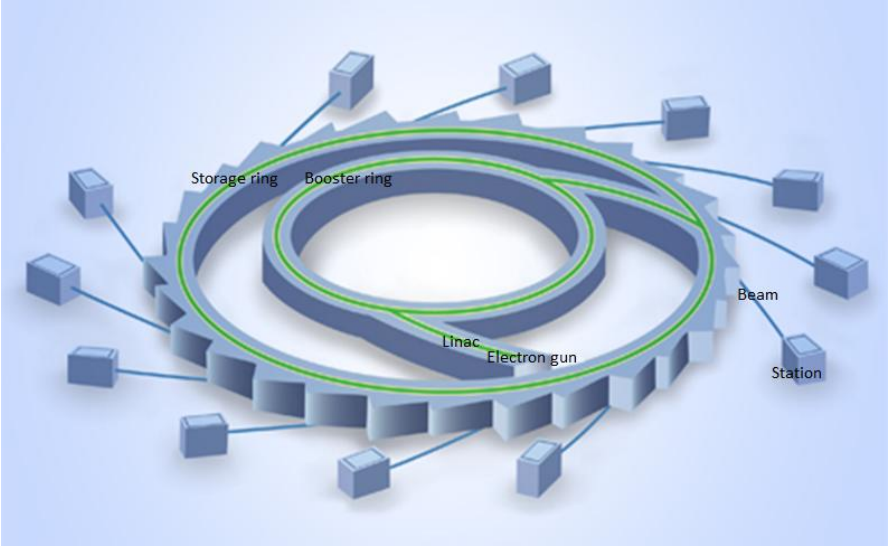

FIGURE 1.10: Schematic diagram of the synchrotron light source.

(Figure 1.10). The synchrotron light source is notable for its high brilliance, high level of polarization, high collimation and wide tunability in wavelength by monochromatization. The high brilliance makes it possible to collect the diffraction pattern of big protein molecules with large unit cells. The wide tunability is also required for anomalous dispersion.

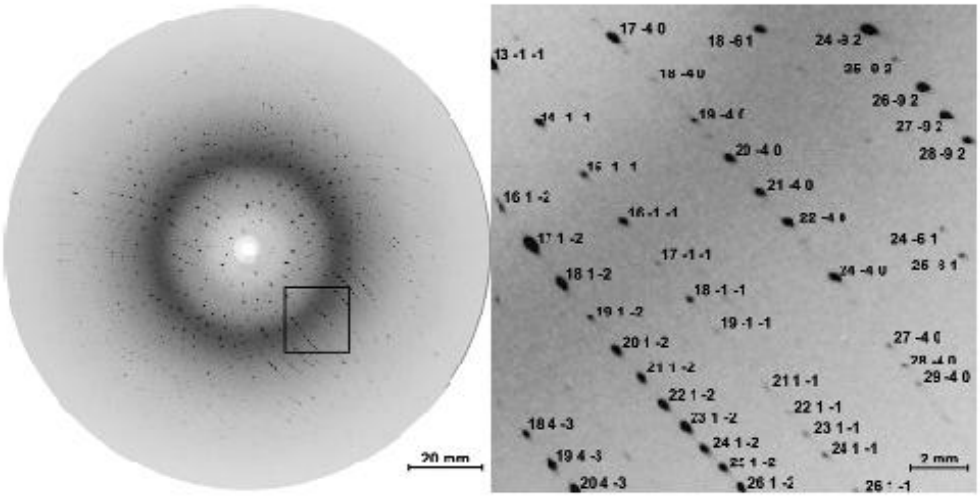

FIGURE 1.11: X-ray diffraction pattern and diffraction-spot indexing.

A typical diffraction pattern contains thousands of diffraction spots. A diffraction pattern and its indexed diffraction spots are shown in Figure 1.11. Each diffraction spot can be referred to as a reflection. The geometrical arrangement of the reflections gives the information about the cell dimensions and the symmetry of the crystal. The intensities of the reflections yield the information about the content in a unit cell. Each diffraction spot corresponds to a reciprocal-lattice point. There are three integers associated with each reciprocal-lattice point. Indexing is to find the cell parameters, the space group and the three integers for each diffraction spot on each diffraction pattern. Some software packages are available for auto indexing.

A space group can be determined from the Laue symmetry and the reflection conditions obtained from the diffraction patterns. Space group determination starts with the assignment of a Laue class to the weighted reciprocal lattice, and the determination of the cell geometry. The conventional cell is selected that the basis vectors can coincide with the highest symmetry directions. The Laue class determines the crystal system. On the diffraction pattern, sets of reflections can be systematically absent. These extinctions imply the presence of a centered cell or the presence of symmetry elements with glide or screw components. Usually Laue class plus reflection conditions can't uniquely determine the space group. In this case, other information such as the presence or absence of an inversion center in the crystal unit cell should be examined to overcome the ambiguities.

1.5 Protein model building and model refinement

From the geometrical arrangement of the reflections on the diffraction pattern, the cell parameters, the space group and the Miller indices of each reflection can be determined. From the intensity of the reflections, the electron density function of the unit cell can be determined by solving the phase problem which will be discussed in Chapter 3.

The process to build an atomic model structure into a calculated electron density map is called model building. A good electron-density map should be interpretable. Secondary structures clearly show up, such as alpha helices and beta sheets. The main chain can be located and traced. An atomic model structure can be fitted into the calculated electron density map by hand or some auto-building software packages, as shown in Figure 1.12.

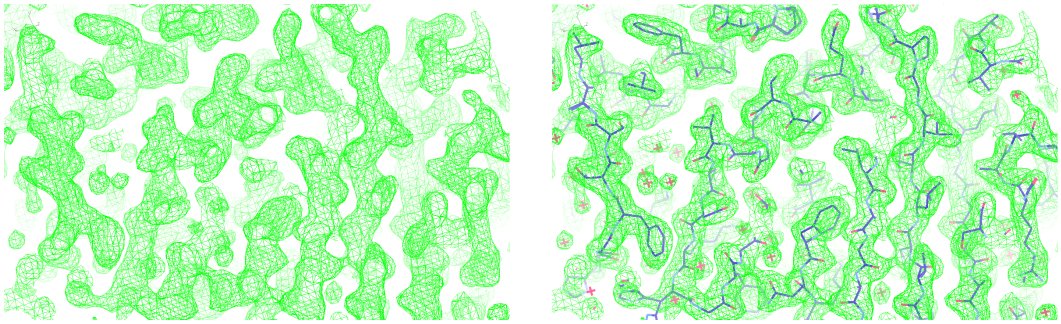

FIGURE 1.12: Calculated electron density map (green) and the built-in model (blue).

The initial model is often not complete or accurate. Iterative adjustments on the model are necessary. New structure factors can be calculated from the model. Model refinement is to minimize the discrepancy between the calculated structure factors of the model and the experimental observed structure factors. Several parameters are defined to monitor the accuracy of the model.

R-factor is defined to measure the difference between the calculated structure factor and the observed structure factor.

$$R = \frac{\sum \big||F_{obs}| - k|F_{cal}|\big|}{\sum |F_{obs}|} \tag{1.4}$$

k is a scale factor. For a random model structure, the value of the R-factor is about 0.59. For a good refined model structure, R should be less than 0.30.

R_{free} and R_{work} are another two parameters used to measure the deviation between the calculated structure factors and the observed structure factors[5]. At the beginning of the phase retrieval, some reflections in the observed data can be randomly selected and are not used for phasing and refinement. Generally, 5% observed reflections can be randomly selected and set aside. These reflections are called free data set or test data set. The remained 95% observed reflections are called working data set and are used to retrieve phases and refine the model. R_{free} is calculated from the free data set. Because the free data set is not used during phase retrieval and model refinement, R_{free} is more convincing to show the accuracy of the model structure. R_{work} is calculated from the working data set. For a random model or random electron density map, both R_{work} and R_{free} are close to 0.59. For an accurate model or correct eletron density map, R_{free} and R_{work} should be less than 0.30.

$$R_{free} = \frac{\sum_{hkl \in free} \big||F_{obs}| - k|F_{cal}|\big|}{\sum_{hkl \in free} |F_{obs}|} \tag{1.5}$$

$$R_{work} = \frac{\sum_{hkl \in work} \big||F_{obs}| - k|F_{cal}|\big|}{\sum_{hkl \in work} |F_{obs}|} \tag{1.6}$$

R factors defined above can only show the overall accuracy of the model structure. The model may contain some local errors and they can't be located by those R factors defined above. In order to find some local errors of the model structure, a new R factor in real space can be defined as follows.

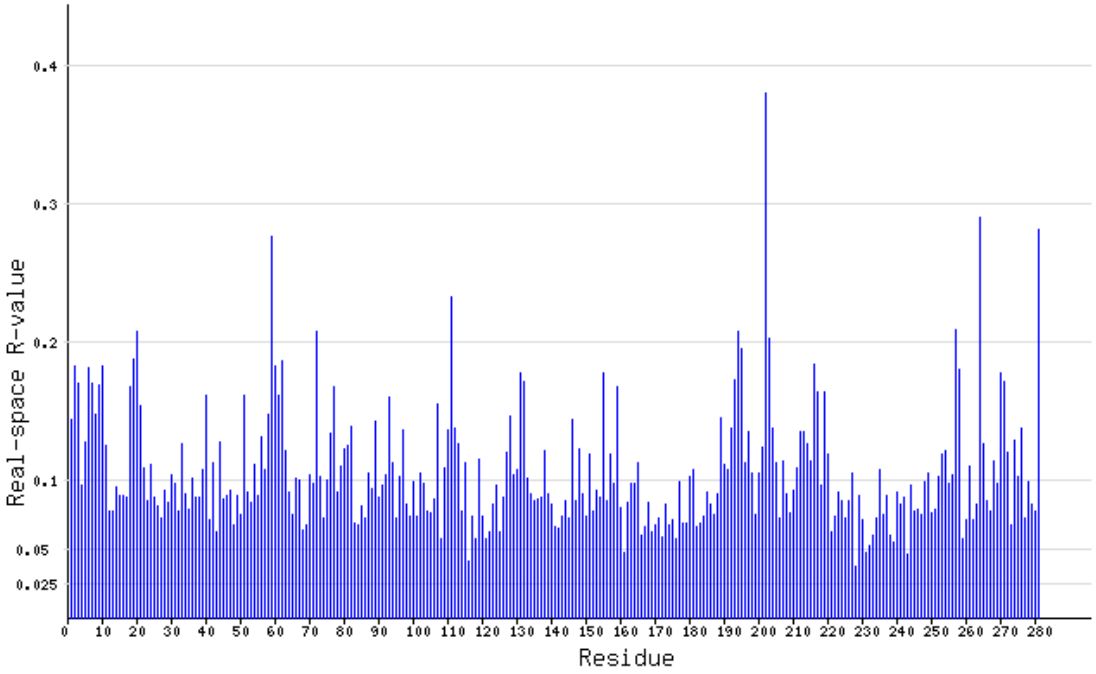

FIGURE 1.13: Real-space R factor for each residue.

$$R_{real} = \frac{\sum|\rho_{obs} - \rho_{cal}|}{\sum|\rho_{obs} + \rho_{cal}|} \tag{1.7}$$

R_{real} is a parameter defined in real space and can be calculated for each amino acid residue (Figure 1.13) [6]. The observed electron density map is the Fourier transform of the observed structure-factor magnitudes combined with the phases calculated from the model. The calculated electron density map is directly computed from the model structure by a Gaussian distribution of electron density around the average position of

each atom in the model.

In addition to R factors, a Ramachandran plot or Ramachandran diagram is usually employed to check the accuracy of the model structure[7]. Ramachandran plot is used to visualize the empirical distribution of the backbone dihedral angles ψ against φ of amino acid residues observed in the model structure in usage for structure validation. The definition of the protein backbone dihedral angles are shown in Figure 1.14. Ramachandran plot shows the possible confirmations of ψ and φ angles for a polypeptide. The angle ω of the peptide bond is usually 180°. The ψ and φ angles in the model structure should fall into the correct regions on the Ramachandran plot. If some angles fall outside of the correct regions, the local structure corresponding to those angles should be adjusted.

The dihedral angle or torsion angle is the angle between two planes. For four consecutively bounded atoms A-B-C-D, atoms A-B-C define the first plane and atoms B-C-D define the second plane. The angle between these two planes is called a dihedral angle.

In Figure 1.14, the dihedral angles on the backbone are defined as follows. The ψ dihedral angle for residue i is defined by N_i–$C\alpha_i$–C_i–N_{i+1}. The φ dihedral angle for residue i is defined by C_{i-1}–N_i–$C\alpha_i$–C_i and the ω dihedral angle for residue i is defined by $C\alpha_{i-1}$–C_{i-1}–N_i–$C\alpha_i$.

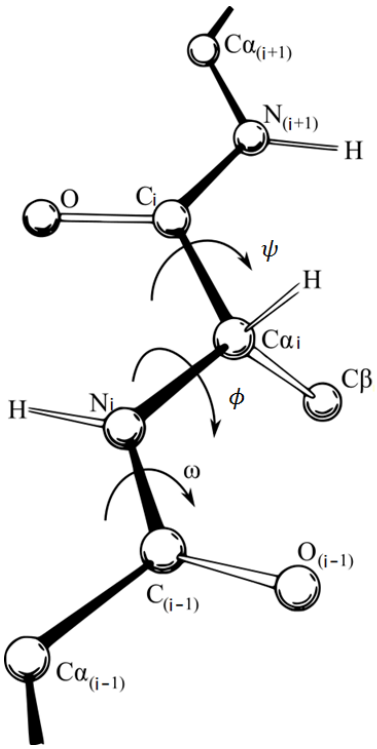

FIGURE 1.14: Protein backbone dihedral angles.

We assume atoms can be treated as hard spheres with van der Waals radii. In Figure 1.15, the white regions are disallowed because they correspond to conformations that atoms in the polypeptide become closer than their van der Waals radii. The red regions are favored and there are no steric clashes. They correspond to conformations such as alpha helices and beta sheets. The yellow regions are allowed, because they correspond to slightly shorter van der Waals radii. Amino acids can't form left-handed helix, but individual residues, such as glycine, occasionally take left-handed conformation.

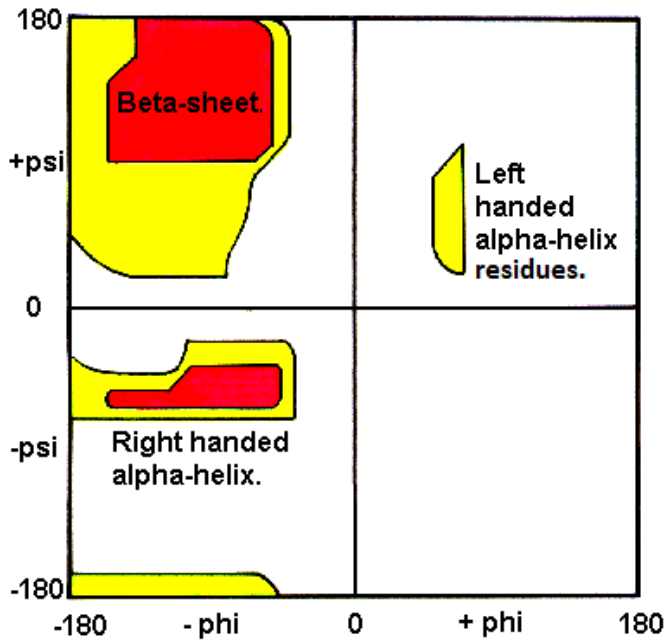

FIGURE 1.15: Ramachandran plot and the distribution of the dihedral angles of alpha helix and beta sheet. The red regions are favored, the yellow regions are allowed and the white regions are disallowed.

During model refinement, if the dihedral angles of a residue fall into the white region, the local conformation of that residue on the model should be adjusted.

1.6 Summary

In this chapter, an introduction about protein X-ray crystallography is described. Protein X-ray crystallography needs a large amount of protein molecules which can be produced by gene cloning and DNA expression. Purified proteins are dissolved into water-buffer solvent and precipitant solution. When the protein solution gets supersaturated, protein nucleation occurs. By changing the solution conditions, the nuclei can grow into big crystals. Protein crystallization is still a big challenge for biochemists. The crystal is mounted to a goniometer head. When the crystal rotates with the head

of the goniometer around three axes, reciprocal-lattice points contacts with the surface of the Ewald sphere and diffracted rays occur. Each rotation of the crystal gives a diffraction pattern. The combination of all diffraction patterns gives the entire diffraction data. The space group and the cell parameters are derived from the geometry of the diffraction pattern. Each diffraction spot on the diffraction pattern is indexed with three Miller indices. The magnitudes of structure factors are computed directly from the intensities of diffraction spots on the diffraction pattern, but the phases are lost. This phase problem is still a challenge for physicists. In Chapters 3, 4, 5 and 6, we will focus on the phase problem.

When the phase problem is solved, the calculated electron density becomes interpretable. A model structure can be built into the calculated electron density map. This model structure is not accurate at the beginning and further refinement is necessary. Several R factors and a Ramachandran plot can be used to monitor the accuracy of the model during the refinement.

Chapter 2

Principles of protein X-ray crystallography

2.1 Introduction

The resolution of a microscopy is limited by the wavelength of the electro-magnetic radiation used. In order to see the atomic structure of proteins, the wavelength of the electron-magnetic radiation used should be around one angstrom which is X-ray.

X-rays are diffracted by electrons in a protein crystal which behaves like a three-dimensional diffraction grating. There are both constructive and destructive interference effects on the diffraction pattern which give discrete diffraction spots known as Bragg reflections. The diffraction pattern is the Fourier transform of the electron density of the protein crystal[1–3, 8–10]. The temperature factor or Debye-Waller factor has to be considered for X-ray diffraction.

In a protein crystal, there are both protein molecules and solvent molecules. The

protein molecules are arranged but the solvent molecules are disordered. The solvent molecules also contribute to the observed diffraction data. Bulk solvent correction is taken into consideration. The synthetic diffraction data with bulk solvent correction is important for better understanding the observed data and testing new phase-retrieval methods.

This chapter is organized as follows. In Section 2, X-ray scattering by electrons is described. In Section 3, X-ray diffraction by a crystal is discussed. In section 4, we talk about the Fourier transform of the diffraction pattern which gives the electron density in the unit cell. In Sections 5 and 6, temperature factor and atomic radius are analyzed. In Section 7, we discuss the calculation of synthetic diffraction data of a protein crystal. The last section is a summary.

2.2 X-ray scattering by electrons

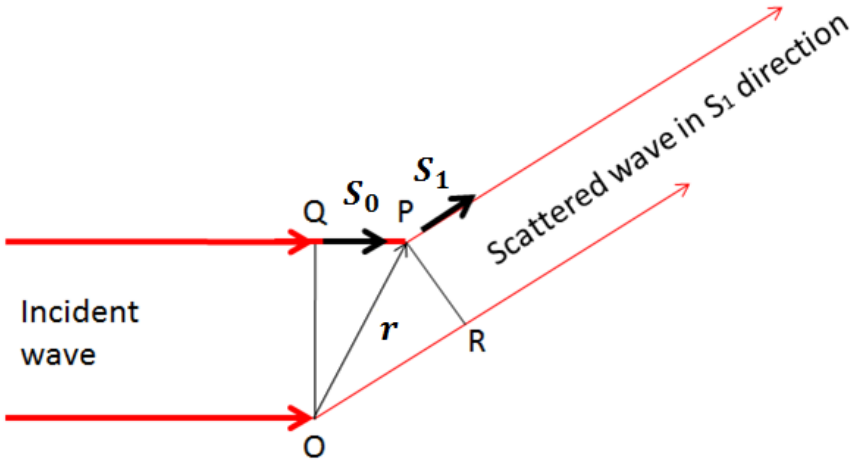

FIGURE 2.1: X-ray scattering by electrons.

Suppose there are two electrons at point O and point P, respectively, as shown in Figure 2.1. The displacement between the two electrons is $\mathbf{r}$. The incident rays, indicated by a wave vector $\mathbf{S_0}$ with a magnitude of $1/\lambda$, are scattered by the two electrons. The lower ray passing along point O follows a longer path than the upper ray which passes point P. The path difference between the two scattered rays in $\mathbf{S_1}$ direction depends on the displacement between the two electrons and the direction of the scattered rays. The path difference can be written as follows.

$$\overline{QP} - \overline{OR} = \mathbf{r} \cdot \mathbf{S_0}\lambda - \mathbf{r} \cdot \mathbf{S_1}\lambda = \mathbf{r} \cdot (\mathbf{S_0} - \mathbf{S_1})\lambda \tag{2.1}$$

With respect to the phase of the lower ray, the phase difference between the scattered rays is

$$-2\pi\frac{\overline{QP} - \overline{OR}}{\lambda} = -2\pi\frac{\mathbf{r} \cdot (\mathbf{S_0} - \mathbf{S_1})\lambda}{\lambda} = 2\pi\mathbf{r} \cdot (\mathbf{S_1} - \mathbf{S_0}) = 2\pi\mathbf{r} \cdot \mathbf{S} \tag{2.2}$$

where $\mathbf{S} = \mathbf{S_1} - \mathbf{S_0}$.

The scattered ray in $\mathbf{S_1}$ direction can be regarded as a reflection on a group of parallel planes perpendicular to vector $\mathbf{S}$.

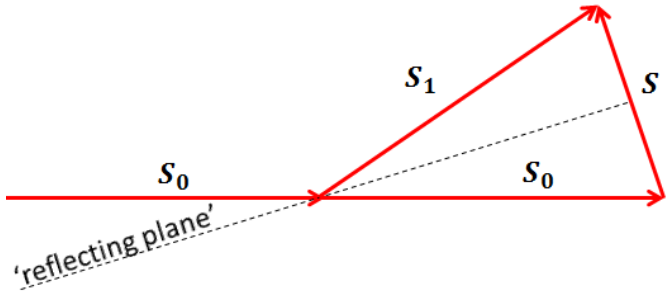

FIGURE 2.2: The scattered ray can be regarded as being reflected against a plane.

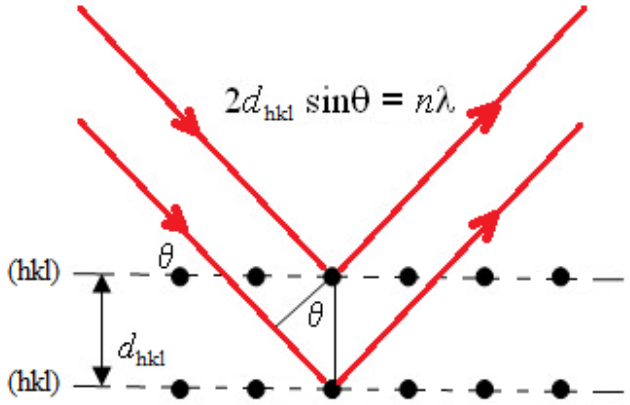

FIGURE 2.3: Bragg's law which is the condition to produce strong diffracted rays.

$$2d_{hkl}\sin\theta = n\lambda \tag{2.3}$$

Equation 2.3 is Bragg's law.

Suppose we make a translation of the origin which is shown in Figure 2.4. The upper scattered ray has phase $2\pi\mathbf{r_2}\cdot\mathbf{S}$ and the lower scattered ray has phase $2\pi\mathbf{r_1}\cdot\mathbf{S}$. If we add the two scattered rays, the sum is as follows.

$$e^{2\pi i\mathbf{r_2}\cdot\mathbf{S}} + e^{2\pi i\mathbf{r_1}\cdot\mathbf{S}} = (e^{2\pi i\mathbf{r}\cdot\mathbf{S}} + 1)e^{2\pi i\mathbf{r_1}\cdot\mathbf{S}} \tag{2.4}$$

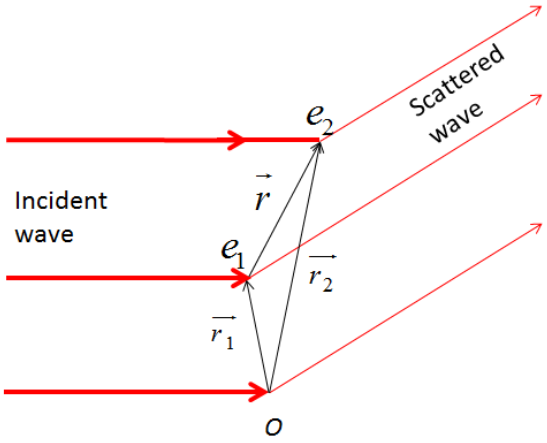

FIGURE 2.4: A shift of the origin causes a shift of the phase.

Translation of the origin results in the same phase difference for all scattered rays. Translation does not affect the magnitude of the scattered rays.

2.3 X-ray diffraction by a crystal

The atomic scattering factor of an atom with an electron density distribution $\rho(\mathbf{r})$ can be written as follows.

$$f(\mathbf{S}) = \int \rho(\mathbf{r}) e^{2\pi i \mathbf{r}\cdot\mathbf{S}} d^3\mathbf{r} \tag{2.5}$$

The origin is located at the center of the nucleus.

In a unit cell, suppose there are N atoms and the j^{th} atom is located at $\mathbf{r_j}$. The scattering factor of the unit cell can be written as a summation of the scattering factors of all atoms.

$$F(\mathbf{S}) = \sum_{j=1}^{N} F_j = \sum_{j=1}^{N} f_j e^{2\pi i \mathbf{r_j}\cdot\mathbf{S}} \tag{2.6}$$

f_j is the atomic scattering factor of the j^{th} atom.

In a crystal, suppose there are $N_1 \times N_2 \times N_3$ unit cells. The scattering factor of other unit cells can be written as the scattering factor of one unit cell times a phase due to the unit-cell translation. The total scattering factor of a crystal can be written as a summation of the scattering factors of all unit cells.

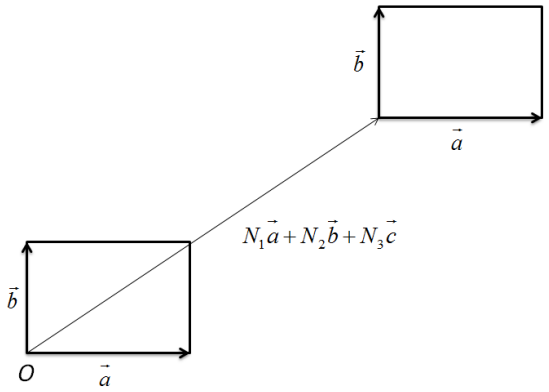

FIGURE 2.5: A crystal contains a great number of identical unit cells. Any two unit cells can be related by a translation.

$$\begin{aligned} K(\mathbf{S}) &= \sum_{q_1=1}^{N_1}\sum_{q_2=1}^{N_2}\sum_{q_3=1}^{N_3} F(\mathbf{S}) e^{2\pi i q_1 \mathbf{a}\cdot\mathbf{S}} e^{2\pi i q_2 \mathbf{b}\cdot\mathbf{S}} e^{2\pi i q_3 \mathbf{c}\cdot\mathbf{S}} \\ &= F(\mathbf{S}) \sum_{q_1=1}^{N_1} e^{2\pi i q_1 \mathbf{a}\cdot\mathbf{S}} \sum_{q_2=1}^{N_2} e^{2\pi i q_2 \mathbf{b}\cdot\mathbf{S}} \sum_{q_3=1}^{N_3} e^{2\pi i q_3 \mathbf{c}\cdot\mathbf{S}} \end{aligned} \tag{2.7}$$

Because the summation is over all unit cells in the crystal, the summation $\sum_{q_1=1}^{N_1} e^{2\pi i q_1 \mathbf{a}\cdot\mathbf{S}}$ and the other two summations over q_2 and q_3 are almost always zero. In order to get a none zero $K(\mathbf{S})$, Laue conditions should be satisfied.

$$\begin{aligned} \mathbf{a}\cdot\mathbf{S} &= h \\ \mathbf{b}\cdot\mathbf{S} &= k \\ \mathbf{c}\cdot\mathbf{S} &= l \end{aligned} \tag{2.8}$$

Therefore, for a large crystal the scattering factor of the crystal is the same as the scattering factor of the unit cell with the satisfaction of the Laue conditions.

2.4 Fourier transform of the diffraction pattern

When a monochromatic X-ray diffracts off a crystal, it performs part of a mathematical operation, the Fourier transform. When the incidence angle is varied by rotating the crystal, the complete transform is produced and the whole diffraction data can be recorded on a set of diffraction patterns. The flaw of this perfect transform is that people can't measure the phase of the diffracted wave. Otherwise the entire protein structure can be computed by an inverse Fourier transform.

If $\rho(\mathbf{r})$ is the electron density function of all atoms in the unit cell, the structure factor of the unit cell can be represented as an integral of $\rho(\mathbf{r})$.

$$F(\mathbf{S}) = \int_{unitcell} \rho(\mathbf{r}) e^{2\pi i \mathbf{r}\cdot\mathbf{S}} dv \tag{2.9}$$

$\mathbf{r} = x\mathbf{a} + y\mathbf{b} + z\mathbf{c}$, where x, y and z are fractional coordinates.

$$\mathbf{r}\cdot\mathbf{S} = (x\mathbf{a} + y\mathbf{b} + z\mathbf{c})\cdot\mathbf{S} = x\mathbf{a}\cdot\mathbf{S} + y\mathbf{b}\cdot\mathbf{S} + z\mathbf{c}\cdot\mathbf{S} = hx + ky + lz \tag{2.10}$$

$$dv = \mathbf{a}dx \cdot \mathbf{b}dy \times \mathbf{c}dz = V dxdydz \tag{2.11}$$

We can transfer the integral from orthogonal coordinates to fractional coordinates.

$$F(h,k,l) = V \int_{x=0}^{1}\int_{y=0}^{1}\int_{z=0}^{1} \rho(x,y,z) e^{2\pi i(hx+by+lz)} dxdydz \tag{2.12}$$

If we divide the unit cell into a grid of regularly spaced points, the density on each grid point is denoted as $\rho_j(x_j, y_j, z_j)$. Then the integral can be written as a summation over all grid points inside the unit cell.

$$F(h,k,l) = \frac{V}{N_x N_y N_z} \sum_{j_1=1}^{N_x} \sum_{j_2=1}^{N_y} \sum_{j_3=1}^{N_z} \rho(x_j, y_j, z_j) e^{2\pi i(hx_j + by_j + lz_j)} \tag{2.13}$$

where $dxdydz = \frac{1}{N_x N_y N_z}$. N_x, N_y and N_z are the number of grid points in the x, y and z directions.

Since the integral format of the structure factor is the Fourier transform of the electron density function in the unit cell, the inverse Fourier transform of the structure factor gives the electron density function. Because of Laue conditions, the reciprocal space is discretized. The integral over reciprocal space should be replaced by a summation over all Miller indices.

$$\rho(x_j, y_j, z_j) = \frac{1}{V} \sum_{h=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} F(h,k,l) e^{-2\pi i(hx_j + by_j + lz_j)} \tag{2.14}$$

Because $F(h,k,l)$ is complex which consists of a magnitude and a phase. The magnitude is proportional to the square root of the diffraction intensity on the diffraction pattern which can be recorded in the experiment. However, the phase is lost.

$$\rho(x_j, y_j, z_j) = \frac{1}{V} \sum_{h=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \left|F(h,k,l)\right| e^{-2\pi i(hx_j + by_j + lz_j) + i\alpha(h,k,l)} \tag{2.15}$$

In order to find the electron density function of the unit cell, the phase for each reflection should be found out first. This is called the phase problem.

2.5 Temperature factor

In the X-ray diffraction experiment, the temperature of the crystal is not zero even the crystal is cooled with nitrogen stream. A temperature factor can blur the electron density function in the unit cell. Therefore, the temperature factor should be considered into the scattering factor of the unit cell. Here we consider isotropic temperature factor. Temperature factor can cause vibrations of the electron density. These vibrations can be decomposed as parallel and perpendicular components to the reflecting plane. The parallel component does not contribute to the scattering factor. But the perpendicular component does affect the scattering factor.

$$F(\mathbf{S}) = \sum_{j=1}^{N} f_j e^{2\pi i \mathbf{r_j} \cdot \mathbf{S}} e^{-\frac{B}{4} S^2} \tag{2.16}$$

The temperature factor B is also known as Debye-Waller factor. B factor is positive and it reduces the magnitude of the structure factor. The temperature factor has a bigger effect on higher-resolution reflections than lower-resolution reflections. But the temperature factor does not affect the phase of the structure factor. It can be shown that the thermal parameter B is related to the mean square displacement $\overline{u^2}$ of the atomic vibration.

$$B = 8\pi^2 \overline{u^2} \tag{2.17}$$

2.6 Atomic radius

Suppose all atoms in the unit cell are the same and each atom can be represented by a Gaussian sphere with density $\rho(\mathbf{r})$.

$$\rho(\mathbf{r}) = \frac{Z}{a_0^3\sqrt{\pi^3}} e^{-\frac{|\mathbf{r}|^2}{a_0^2}} \tag{2.18}$$

where Z is the atomic number and a_0 is the radius of the Gaussian sphere. The atomic scattering factor is the Fourier transform of the Gaussian function which gives another Gaussian function in reciprocal space.

$$f(\mathbf{S}) = Ze^{-\pi^2 a_0^2 S^2} \tag{2.19}$$

Then the structure factor of the unit cell can be written as $F(\mathbf{S}) = \sum_{j=1}^{N} f_j e^{2\pi i \mathbf{r_j}\cdot\mathbf{S}}$. If the atomic radius increases from a_0 to a_P , then the new structure factor becomes

$$G(\mathbf{S}) = F(\mathbf{S})e^{-\pi^2(a_P^2 - a_0^2)S^2} \tag{2.20}$$

Increasing atomic radius has a similar effect as increasing the temperature factor. Both affect only the magnitude of the structure factor. Higher-resolution structure factors can be attenuated much more than the lower-resolution structure factors. The phase of the structure factor does not change.

2.7 Synthetic diffraction data of a protein crystal

After model building, synthetic diffraction data should be calculated during the refinement. Synthetic data includes the diffraction contributions from the protein molecules and the solvent molecules. The atomic model of a protein often includes thousands of atoms. Different kinds of atoms have different atomic scattering factors. Each atom should have its own temperature factor. The electron density in the bulk solvent region is basically treated as a constant. A solvent mask has to be defined for calculating the diffraction contribution of the bulk solvent[11].

The total structure factor used in refinement can be written as follows.

$$F_{\text{total}} = k_{\text{overall}} e^{-SU_{\text{crystal}}S^T} (F_{\text{protein model}} + F_{\text{bulk solvent}}) \tag{2.21}$$

In our calculation, for simplicity, we suppose k_{overall} equals 1, and ignore the anisotropy factor $e^{-SU_{\text{crystal}}S^T}$. $F_{\text{protein model}}$ is the structure factor calculated from the protein model. It includes the contributions of all atoms in the model. $F_{\text{bulk solvent}}$ is the structure factor contributed by the bulk solvent.

$F_{\text{protein model}}$ is the structure factor calculated from the protein model which contains thousands of atoms. Different atoms have different scattering factors. Each atom should have a temperature factor. For simplicity, a single Gaussian function can be used as the atomic scattering factor. However, in practice, the atomic scattering factor is much more complicated than a single Gaussian function. A linear combination of several Gaussian functions should be used as the atomic scattering factor. Generally, the atomic analytical scattering factor can be expressed as a linear combination of five Gaussian functions and

a constant term[12].

$$f(S) = \sum_{i=1}^{5} a_i e^{-b_i S^2} + c \tag{2.22}$$

where a_i and b_i are parameters specific to atom type.

The protein model contains all the non-hydrogen atoms. The sum of all scattering factors with proper phases over all atoms in the unit cell gives the calculated structure factor of the protein model denoted as $F_{\text{protein model}}$. In practice, this summation can be written as a sum over all atoms in the asymmetric unit and their corresponding equivalent-position atoms in the unit cell.

$$F_{\text{protein model}}(\mathbf{S}) = \sum_{j=1}^{N} q_j f_j e^{2\pi i \mathbf{r_j} \cdot \mathbf{S}} e^{-\frac{B_j}{4} S^2} \tag{2.23}$$

where q_j , B_j, and $r_j = (x_j, y_j, z_j)$ are atomic occupancy, isotropic temperature factor, and fractional coordinates of the j^{th} atom. The occupation number is the fraction of unit cells that contain the atom in this particular location. The temperature factor has to be considered in the atomic structure factor. It can be different for different atoms. Generally, atoms buried in the protein have smaller temperature factors. Atoms near the protein surface often have big temperature factors. These atoms can move a little bit around their equilibrium positions.

In addition to the protein model, bulk solvent correction should also be included into the synthetic data calculation. In the solvent region of the protein crystal, the solvent molecules are disordered. This is an important difference between the traditional crystal and the protein crystal. In traditional crystal, all atoms have the same orientation in different crystal unit cells. However, in protein crystal, only protein molecules and

some attached solvent molecules have the same orientation in different crystal unit cells. Therefore, an average electron density in the solvent region can be employed to stand for the solvent molecules.

$$F_{\text{bulk solvent}} = k_{\text{sol}} e^{-\frac{B_{\text{sol}} S^2}{4}} F_{\text{solvent mask}} \tag{2.24}$$

$F_{\text{solvent mask}}$ is the structure factor calculated from the solvent mask with a unit density. k_{sol} is the average solvent density. B_{sol} is the average temperature factor of the solvent.

Before an average electron density is used to represent the solvent molecules, the exact solvent region should be identified inside the unit cell. Generally, solvent molecules are water molecules. Each water molecule can be approximately treated as a sphere with a radius r_{solvent} or r_{probe}. This radius is used to find the solvent-accessible surface which is a surface accessible to the solvent molecules. The solvent-accessible surface is typically designed by the 'rolling ball' method. This method uses a sphere of solvent with a particular radius to probe the surface of the protein molecule, as shown in Figure 2.6.

Solvent probe is a sphere that approximates the effective size of the solvent molecule. Solvent probe rolls over the model molecular surface. Path of the center of the solvent sphere gives the solvent-accessible surface. It is larger (more external) than the protein surface. Protein surface is defined by all atoms with van der Waals radius corresponding to atom type. After we find the solvent-accessible surface, a contact and reentrant surface can be defined. Any grid points between the accessible surface and the protein molecule should be tested. If the distance between this grid point and the nearest

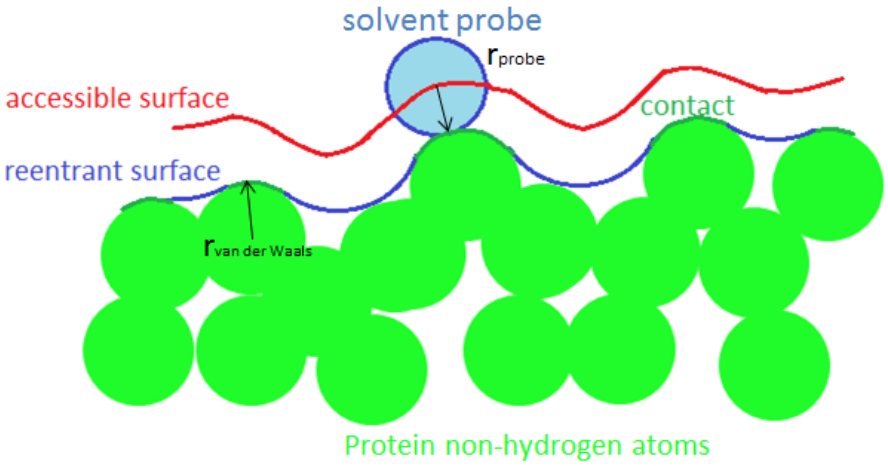

FIGURE 2.6: Accessible surface of the protein molecule.

accessible surface is less than a shrink radius, this grid point will be filled with solvent.

In practice, the solvent mask is defined as follows. The unit cell is divided into a grid of regularly spaced points. A map is defined inside an asymmetric unit and the value of the map is restricted to zero and one. Grid points on the map have an initial value one. All grid points on the map within a distance of $r_{\text{probe}} + r_{\text{rvan der Waals}}$ from atom i is set to zero. r_{probe} is the radius of the solvent molecule. $r_{\text{rvan der Waals}}$ is the van der Waals radius of the protein atom. Those grid points with value one defines the solvent-accessible region. A shrink radius is employed to extend the solvent access region. All grid point marked zero is tested to see if there is a grid point marked one within a distance r_{shrink}. If this happens, the tested grid point is set to one. After solvent extension, the bulk solvent is in close contact with the surface of the protein molecule. All grid points marked one defined the solvent mask which can be used to calculate $F_{\text{solvent mask}}$.

The probe radius or the solvent radius is about 1.4Å. The shrink radius is about 0.8Å. Sometimes, we adopt $r_{\text{probe}} = 1.11$Å and $r_{\text{shrink}} = 0.9$Å[13]. In order to find a pair

of good numbers for r_{probe} and r_{shrink}, one can check the volume of the solvent mask. When the volume of the solvent mask is several percent less than the calculated solvent volume of the crystal, better results can be obtained.

On the solvent mask, all grid points are set to one. Grid points outside the solvent mask are set to zero. Because the electron density of the solvent should be less than one, a scale factor k_{sol} is used to scale the electron density on the solvent mask. k_{sol} represents the average electron density of the solvent (or crystallization buffer) that ranges from $0.33e/\mathring{A}^3$ for pure water to $0.41e/\mathring{A}^3$ for 4M ammonium sulphate. In experiment, the solvent always has a certain finite temperature. A temperature factor B_{sol} is employed to smooth the electron density between the solvent mask and the protein surface. The temperature factor B_{sol} can be from $15\mathring{A}^2$ to $200\mathring{A}^2$.

k_{sol} and B_{sol} can be determined by progressively minimizing a target function $G(p,k)$[11]. $G(p,k)$ measures the difference between the calculated structure factor and the observed structure factor.

$$G(p,k) = \frac{\sum_{h,k,l}(|F_{\text{obs}}| - k_{\text{scale}}|F_{\text{total}}|)^2}{\sum_{h,k,l}|F_{\text{obs}}|^2} \tag{2.25}$$

$$k_{\text{scale}} = \frac{\sum_{h,k,l}|F_{\text{obs}}||F_{\text{total}}|}{\sum_{h,k,l}|F_{\text{obs}}|^2} \tag{2.26}$$

Typically, the initial values of $k_{\text{sol}} = 0.40e/\mathring{A}^3$ and $B_{\text{sol}} = 200\mathring{A}^2$ are chosen. When B_{sol} is being refined, k_{sol} is fixed. When k_{sol} is being refined, B_{sol} is fixed. This process is repeated until a pair of k_{sol} and B_{sol} are reached with a minimum value of $G(p,k)$.

After the optimized k_{sol} and B_{sol} are found, $F_{\text{bulk solvent}}$ is calculated. The total synthetic structure factor is the sum of $F_{\text{bulk solvent}}$ and $F_{\text{protein model}}$. The bulk solvent

correction is very important. Take the observed data of a protien structure as an example. The protein is a formaldehyde-activating enzyme (Fae) with PDB ID 1Y5Y[14] R value defined by Equation 1.4 is calculated to show the difference between the magnitudes of the calculated and the observed structure factors. Without bulk solvent correction, the calculated R value is very big, especially for low resolution shells. However, after the bulk solvent correction, R value is dramatically reduced, because the bulk solvent has an average electron density which obviously contributes to the observed magnitudes.

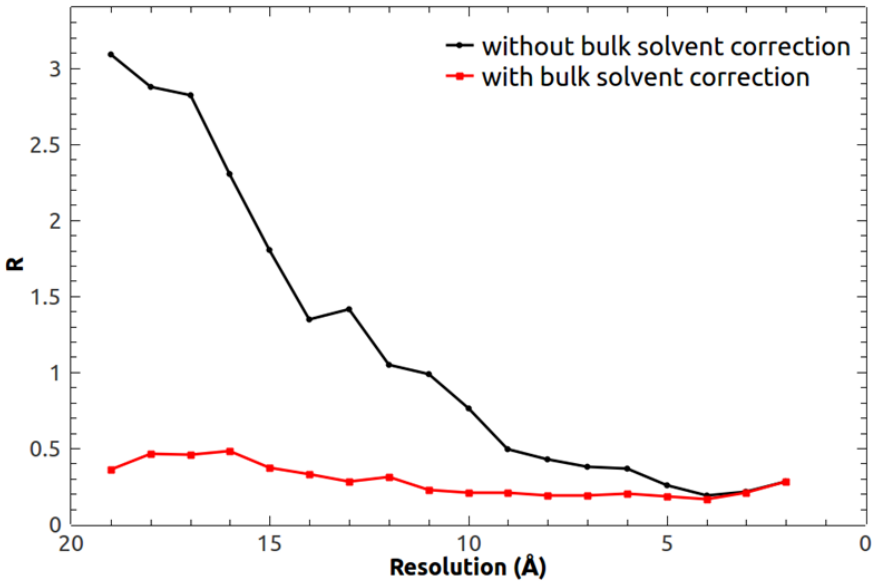

FIGURE 2.7: R values before and after bulk solvent correction.

2.8 Summary

In this chapter we have introduced the principles of protein X-ray diffraction. We have discussed X-ray scattering by electrons, X-ray diffraction by a crystal, the Fourier transform between the structure factor and the electron density, the temperature factor, and

the atomic radius. For a real protein crystal, a bulk solvent correction must be considered, because the solvent region has a constant density which obviously contributes to the observed diffraction data.

Chapter 3

Methods for solving the phase problem

3.1 Introduction

In order to find the protein structure, the phase problem needs to be solved first. There are various means to retrieve the phases.

Heavy-atom method is often used in experimental phasing, such as single isomorphous replacement (SIR), multiple isomorphous replacement (MIR), single anomalous dispersion (SAD) and multiple anomalous dispersion (MAD). For these methods, a native protein crystal and its heavy-atom derivative crystal should be prepared. The difference between the diffraction patterns of the native protein crystal and its heavy-atom derivative crystal gives the diffraction contribution of the heavy atoms alone. Heavy atoms are located by analyzing Patterson maps. Structure factors of the heavy atoms

alone is calculated. Since the structure factor of the heavy-atom derivative is the algebraic sum of the structure factors of the native protein and the heavy atoms, the phase of the structure factor of the native protein crystal is derived by solving some algebraic equations.

Molecular replacement is another method used to find the structure of a target molecule. The amino acid sequence of the target molecule is available before structure determination. Molecules with similar amino acid sequences are selected from the Protein Data Bank. These molecules serve as models of the target molecule. The model molecule is assigned a proper orientation and location in the asymmetric unit in order to reach a maximum overlap with the target molecule. The phases calculated from the model molecule are good estimates for the target molecule.

In addition to the heavy-atom method and the molecular replacement method, there are some direct phasing methods[15–26]. Direct method tries to retrieve the phase directly from the observed data. Different constraints in real space and reciprocal space have been used by direct methods. According to the constraints, direct methods are classified as real-space, reciprocal-space and dual-space direct methods. Generally, the dual-space direct method has better performance.

In all cases, the obtained phases are good estimates of the true phases and they are improved by standard phase-improvement techniques.

3.2 Patterson function

The electron density function in the unit cell is the Fourier transform of the structure factors. The magnitude of the structure factor is proportional to the intensity of the

reflection on the diffraction pattern. Patterson function is the Fourier transform of the reflection intensity[27].

$$P(u,v,w) = \frac{1}{V} \sum_{h=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} |F(h,k,l)|^2 e^{-2\pi i(hu+kv+lw)} \tag{3.1}$$

It has been proved that Patterson function can be alternatively written as follows.

$$P(\mathbf{u}) = \int_{\mathbf{r}} \rho(\mathbf{r})\rho(\mathbf{r}+\mathbf{u})dv \tag{3.2}$$

When both $\rho(\mathbf{r})$ and $\rho(\mathbf{r}+\mathbf{u})$ are nonzero, $P(\mathbf{u})$ will show a peak on the Pattern map. Pattern function gives a map of the vectors between atoms. A Patterson peak corresponds to the displacement between two atoms.

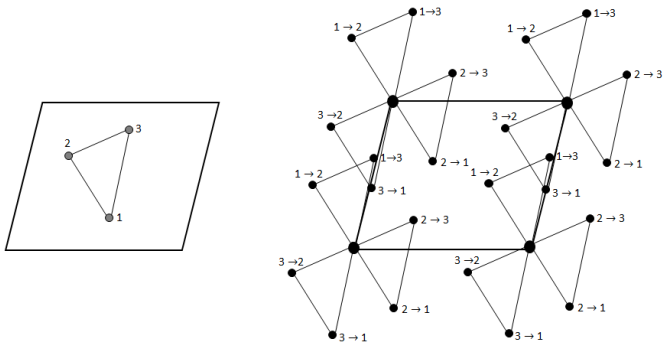

FIGURE 3.1: Structure of a unit cell containing three atoms and construction of the Patterson map.

If there are only a few atoms inside the unit cell,it is possible to work out the locations of the atoms in the unit cell that give the observed Patterson peaks. But if there are hundreds or thousands of atoms in the unit cell, it becomes impossible to deconvolute the Patterson map. However, Patterson function is still very useful for other methods such as isomorphous replacement and anomalous dispersion.

3.3 Isomorphous replacement method

Heavy atoms have more electrons which obviously affect the magnitudes of reflections. Some small molecules already have big atoms such as sulfur. For protein molecules, in order to get strong perturbations on the diffraction pattern, people usually add heavy atoms to the molecule. Addition one or more heavy atoms to a protein molecule is called isomorphous replacement.

Isomorphous replacement method needs at least two diffraction patterns. One is from the native crystal without the addition of heavy atoms. The other is from the derivative crystal which contains heavy atoms and the native molecular structure. The native protein crystal is soaked into a heavy-atom solution, such as solutions with mercury, lead, or gold, which gives a derivative crystal with some heavy atoms attached to the native protein structure. The introduction of heavy atoms should not change the dimensions of the unit cell. The derivative crystal must be isomorphous with the native protein crystal. At least two and often more heavy-atom derivatives are required in isomorphous replacement method.

Heavy atoms of the derivative crystals should be located first. Diffraction pattern of the native crystal gives the structure factor of the protein, $|\mathbf{F}_{\mathrm{P}}|$. Diffraction pattern of the heavy-atom drivative crystal gives the structure factor of the protein plus heavy atoms, $|\mathbf{F}_{\mathrm{PH}}|$. The difference between the two diffraction patterns gives the contribution of heavy atoms alone. The diffraction pattern of the heavy atoms alone is $|\mathbf{F}_{\mathrm{PH}}|-|\mathbf{F}_{\mathrm{P}}|$. A difference Patterson function is constructed, which is the Fourier transform of $|\mathbf{\Delta F}|^2 =$

$(|\mathbf{F}_{\mathrm{PH}}| - |\mathbf{F}_{\mathrm{P}}|)^2$.

$$\Delta P(u,v,w) = \frac{1}{V} \sum_{h=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} |\Delta F(h,k,l)|^2 e^{-2\pi i(hu+kv+lw)} \tag{3.3}$$

Because the difference Patterson function corresponds to the heavy atoms alone, the locations of the heavy atoms are found by analyzing the peaks on the difference Patterson map. Once the locations of heavy atoms are figured out, the structure factors of the heavy atoms $\mathbf{F}_{\mathrm{H}}$ are computed. $\mathbf{F}_{\mathrm{H}}$ corresponds to the structure factor of a unit cell which contains heavy atoms only.

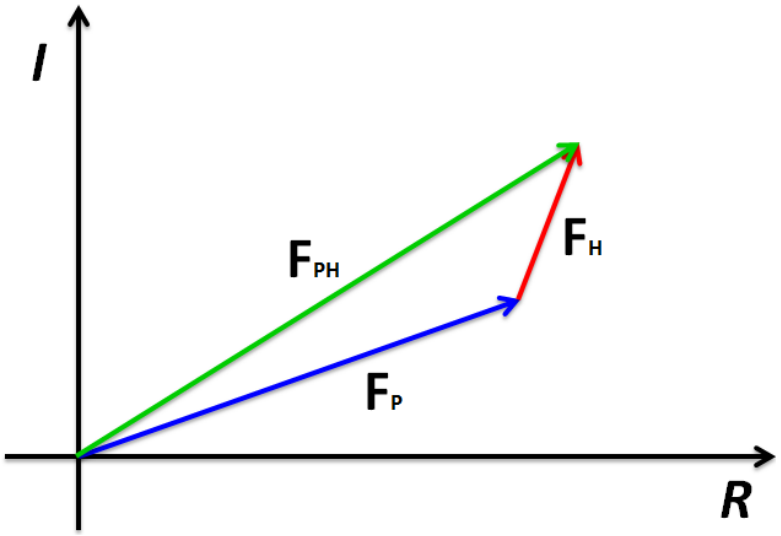

FIGURE 3.2: A structure factor for the heavy atom derivative is the sum of contributions from the native structure and the heavy atom.

The structure factor is a complex number and is represented as a vector on the complex plane. The structure factor of the native crystal is denoted as $\mathbf{F}_{\mathrm{P}}$. The structure factor of the heavy-atom derivative is denoted as $\mathbf{F}_{\mathrm{PH}}$. The relationship between $\mathbf{F}_{\mathrm{P}}$ and $\mathbf{F}_{\mathrm{PH}}$ is written as

$$\mathbf{F}_{\mathrm{P}} = \mathbf{F}_{\mathrm{PH}} - \mathbf{F}_{\mathrm{H}} \tag{3.4}$$

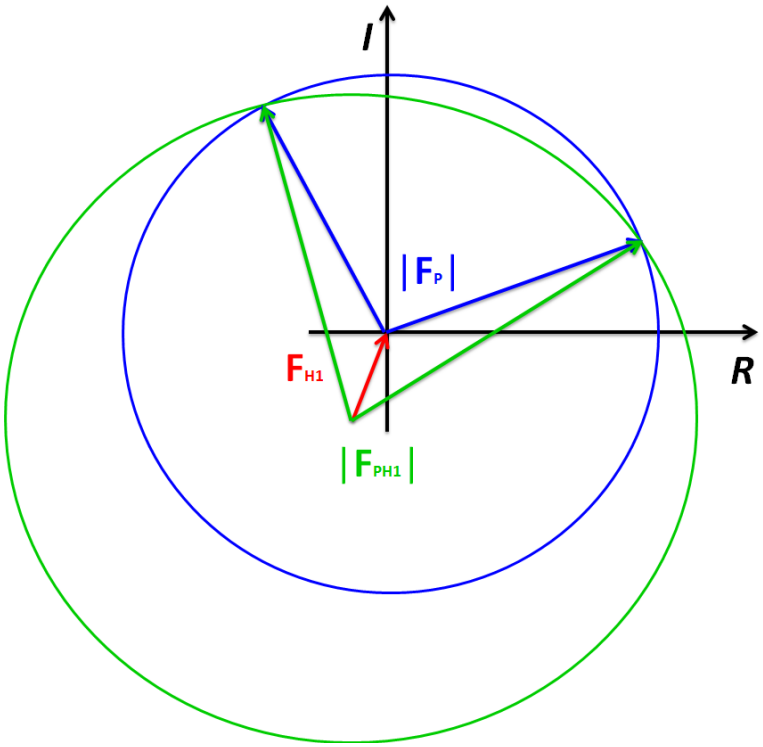

FIGURE 3.3: One heavy-atom derivative indicates two possible phases for the structure factor of the native protein

This relationship is illustrated on the complex plane with the Harker construction or Harker diagram which is shown as two circles in Figure 3.3. One is centered at the origion with a radius $|\mathbf{F}_{\mathrm{P}}|$. The other is centered at the head of $-\mathbf{F}_{\mathrm{H}}$ with a radius $|\mathbf{F}_{\mathrm{PH}}|$. From the Harker construction, two possible solutions of $\mathbf{F}_{\mathrm{P}}$ are found. There are two possible phases for each structure factor of the native crysal. If the two phases are very close, the average of the two phases can be used as an estimate. With the help of phase improvement methods, all phases of the structure factors of the native crysal may

be solved by only one derivative. In this case, the native structure is said to be solved by a single isomorphous replacement (SIR).

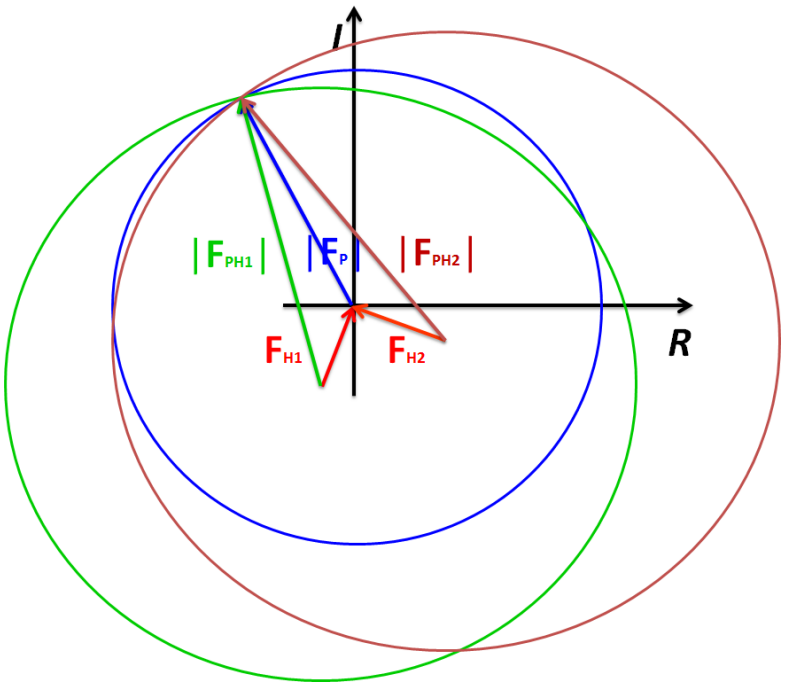

FIGURE 3.4: Two heavy-atom derivatives can fix the phase of the structure factor of the native protein.

Most of the time, the two solutions of $\mathbf{F}_{\mathrm{P}}$ are not close. In order to remove the ambiguity in phases, a second derivative crystal is needed. Using the second derivative, another two possible phases are found for each structure factor of the native crysal, and one of the two phases should be close to one of the previous phases got from the first derivative crystal. In this case, the phase of each structure factor of the native crysal is uniquely determined by two derivative crystals and we call it multiple isomorphous replacement (MIR). Sometimes, in order to completely remove the ambiguity of the phase, more than two derivative crystals are employed.

3.4 Anomalous dispersion method

Heavy atoms absorb X-rays of special wavelength and re-emit X-rays with altered phase, which breaks the Friedel's law. $F_{h,k,l}$ and $F_{-h,-k,-l}$ will not have opposite phases and their magnitudes may be different. This is called anomalous dispersion (AD) or anomalous scattering. If one crystal with anomalous scatters is used to solve the structure, it is called single anomalous dispersion (SAD). If more crystals with different anomalous scatters are used, it is called multiple anomalous dispersion (MAD).

Anomalous scattering is obvious when the wavelength of the X-ray is close to the characteristic emission wavelength of the element. The absorption edges of light atoms, such as carbon oxygen and nitrogen, are far away from the wavelength of X-rays used in diffraction experiment. Only heavy atoms contribute to obvious anomalous scattering on the diffraction pattern. For synchrotron light source, the wavelength of X-rays is tunable, so the abortion edge of different heavy atoms can be reached. If anomalous

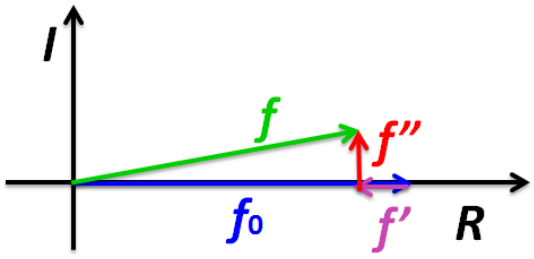

FIGURE 3.5: The anomalous scattering term alters the magnitude and the phase of the atomic scattering factor

dispersion occurs, the total atomic scattering factor consists of three terms.

$$f(\lambda) = f_0 + f'(\lambda) + if''(\lambda) \tag{3.5}$$

f_0 is the normal term which is independent of the wavelength. f' and f'' are anomalous scattering factors which depend on the wavelength. f' is usually negative and f'' is usually positive. f' and f'' are negligible when the wavelength is long or short. The

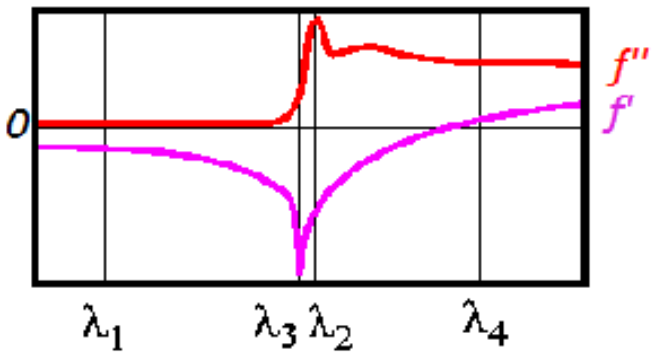

FIGURE 3.6: Schematic experimental values of f' and f'' as a function of X-ray wavelength.

absorption of X-rays drops suddenly at wavelength λ_3, just below the characteristic emission wavelength of the element λ_2. This change in absorption is called absorption edge. At λ_2, f'' has the maximum value. At λ_3, f' reaches its minimum value. Because f'' is responsible for anomalous dispersion, we can tune the wavelength to λ_2 to get a maximal anomalous signal.

Since the diffractive contributions of atoms are additive vectors, the structure factor for the heavy-atom derivative is the vector sum of the structure factors of the protein alone and the heavy atoms alone. $\mathbf{F'_H}$ has been absorbed into $\mathbf{F_H}$. $\mathbf{F''_H}$ corresponds to the heavy-atom anomalous dispersion, which is perpendicular to the normal structure factor $\mathbf{F_H}$. $\mathbf{F_P}$ and $\mathbf{F_H}$ are independent of the wavelength, while $\mathbf{F''_H}$ depends on the wavelength.

$$\mathbf{F_{PH,\lambda}} = \mathbf{F_P} + \mathbf{F_H} + \mathbf{F''_{H,\lambda}} \tag{3.6}$$

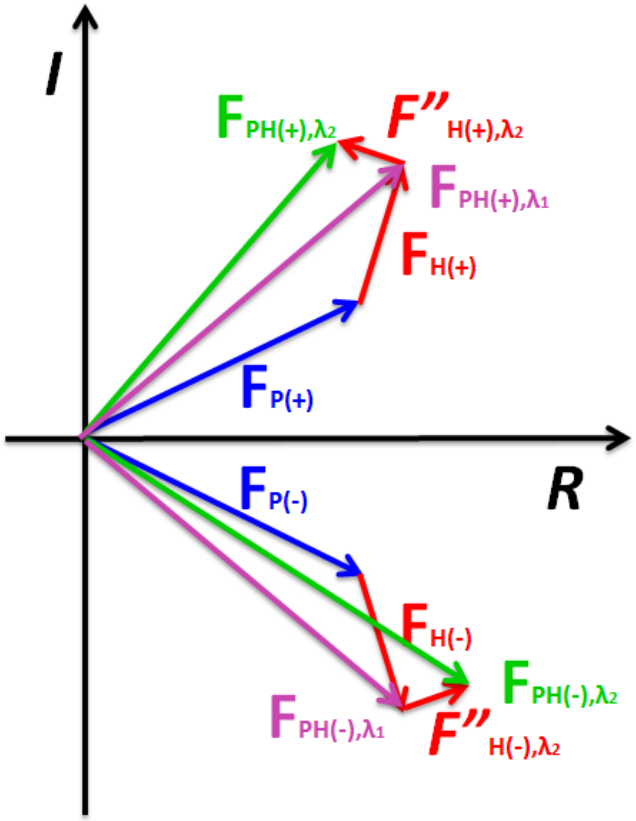

FIGURE 3.7: The structure factor of the anomalous-dispersion derivative breaks Friedel's law under anomalous scattering.

When the wavelength is at λ_1 , which is far away from the absorption edge, heavy atoms have only normal scattering. The structure factor of the heavy-atom derivative $\mathbf{F_{PH,\lambda_1}}$ follows Friedel's law. When the wavelength is at λ_2 , which is at the absorption edge, heavy atoms have strong anomalous scattering. The structure factor of the heavy-atom derivative does not follow Friedel's law. Subscripts $(+)$ and $(-)$ are used to identify positive Miller indices and negative Miller indices.

At wavelength λ_1 , there is no anomalous dispersion. Anomalous dispersion appears at wavelength λ_2. The relationship between $\mathbf{F_{PH(+),\lambda_1}}$ and $\mathbf{F_{PH(+),\lambda_2}}$ is very clear in

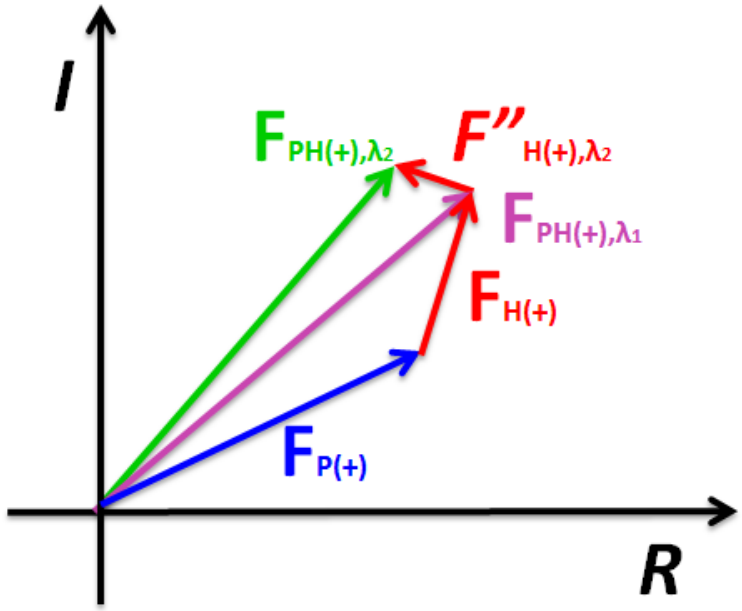

FIGURE 3.8: Vector relationship for Equation 3.7.

Figure 3.8.

$$\mathbf{F_{PH(+),\lambda_1}} = \mathbf{F_{PH(+),\lambda_2}} - \mathbf{F''_{H(+),\lambda_2}} \tag{3.7}$$

The anomalous scattering contribution can be treated as a constant for a given element and roughly independent of reflections. The magnitude of $\mathbf{F''_H}$ is known. The phase of $\mathbf{F''_H}$ depends on the positions of heavy atoms in the unit cell. The locations of heavy atoms can be solved from the peaks on the difference Pattern map obtained from the diffraction pattern of the native protein crystal and the diffraction pattern of the heavy-atom derivative crystal at wavelength λ_1 . So the phase of $\mathbf{F''_H}$ is known.

The vector solution of Equation 3.7 can be found from the Harker diagram in Figure 3.9. We place the vector $-\mathbf{F''_{H(+),\lambda_2}}$ at the origin and draw a circle of radius $|\mathbf{F_{PH(+),\lambda_2}}|$ centered at the head of vector $-\mathbf{F''_{H(+),\lambda_2}}$. All vectors on this circle equal $\mathbf{F_{PH(+),\lambda_2}} -$

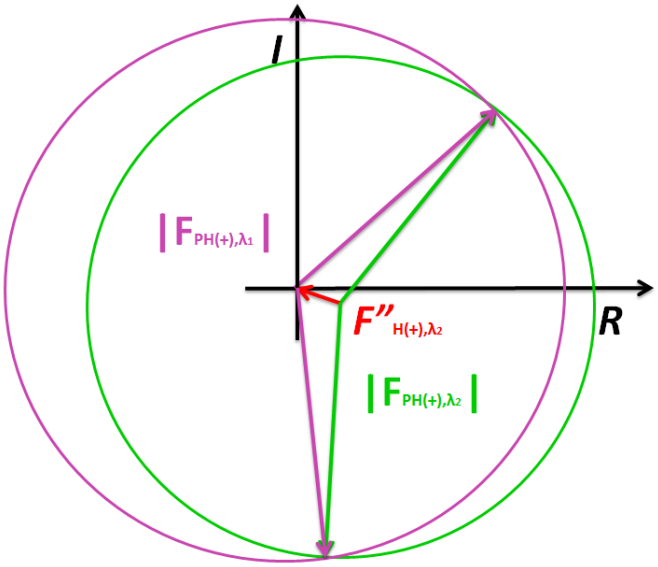

FIGURE 3.9: Vector solution of Equation 3.7. Two possible phases are indicated.

$\mathbf{F''_{H(+),\lambda_2}}$. The head of $\mathbf{F_{PH(+),\lambda_1}}$ lies somewhere on this circle. Then we add another circle of radius $|\mathbf{F_{PH(+),\lambda_1}}|$ centered at the origin. The intersections of the two circles give the two possible solutions of $\mathbf{F_{PH(+),\lambda_1}}$.

From the anomalous dispersion diagram in Figure 3.10, another equation is constructed as follows.

$$\mathbf{F_{PH(+),\lambda_1}} = \mathbf{F^*_{PH(-),\lambda_2}} - (-\mathbf{F''_{H(+),\lambda_2}}) \tag{3.8}$$

The mirror image of $\mathbf{F_{PH(-),\lambda_2}}$ gives the conjugate vector $\mathbf{F^*_{PH(-),\lambda_2}}$ which has the same magnitude with $\mathbf{F_{PH(-),\lambda_2}}$. The mirror image of $\mathbf{F''_{H(-),\lambda_2}}$ gives the conjugate vector $\mathbf{F''^*_{H(-),\lambda_2}}$ which equals $-\mathbf{F''_{H(+),\lambda_2}}$. The three vectors $\mathbf{F_{PH(+),\lambda_1}}$, $-\mathbf{F''_{H(+),\lambda_2}}$, and $\mathbf{F^*_{PH(-),\lambda_2}}$ form a triangle.

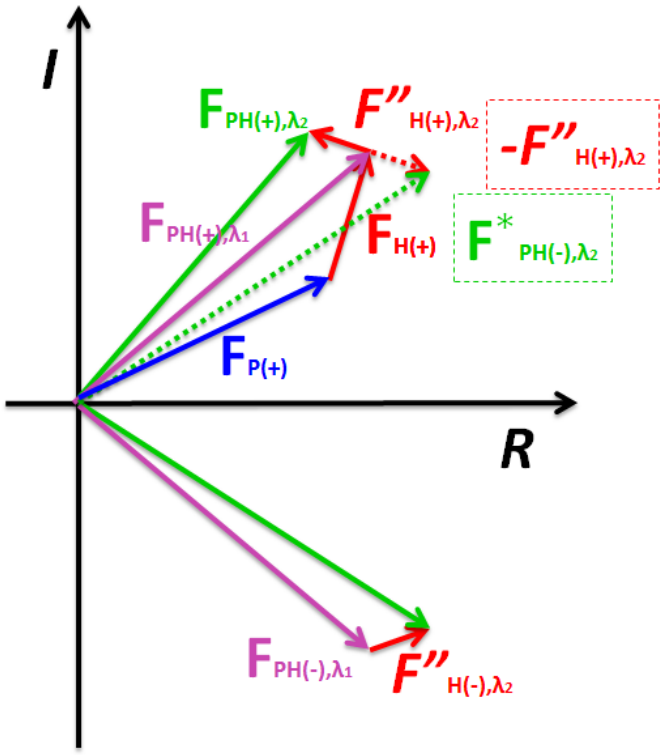

FIGURE 3.10: Vector relationship for Equation 3.8.

In Figure 3.11, a diagram has been drawn to solve Equation 3.8. We place the vector $\mathbf{F''_{H(+),\lambda_2}}$ at the origin and draw a circle of radius $|\mathbf{F_{PH(-),\lambda_2}}|$ centered at the head of vector $\mathbf{F''_{H(+),\lambda_2}}$. All vectors on this circle equal $\mathbf{F^*_{PH(-),\lambda_2}} - (-\mathbf{F''_{H(+),\lambda_2}})$. The head of $\mathbf{F_{PH(+),\lambda_1}}$ lies somewhere on this circle. Then we add another circle of radius $|\mathbf{F_{PH(+),\lambda_1}}|$ centered at the origin. The intersections of the two circles give the another two possible solutions of $\mathbf{F_{PH(+),\lambda_1}}$. Combined with the two possible soluitons got previously, the ambiguity in the phase of $\mathbf{F_{PH(+),\lambda_1}}$ is removed.

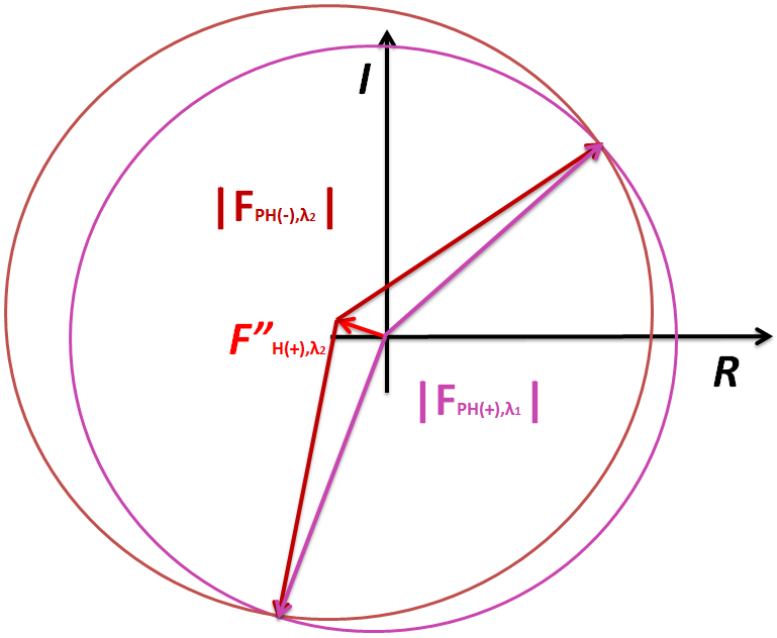

FIGURE 3.11: Vector solution of Equation 3.8. Another two possible phases are indicated. The ambiguity of the phase is removed.

3.5 Molecular replacement method

A known molecular structure is used as a model to estimate the phases of the structure factors of a target molecule when the two molecular structures are similar. This is called molecular replacement[28, 29]. It is the most preferable phasing method, because only the native protein crystal is needed for the X-ray diffraction experiment.

The amino acid sequence of the target protein molecule is available after protein expression. Structures with similar amino acid sequences can be found by searching Protein Data Bank. The same symmetry operations and cell parameters of the target structure should be used for the model structure.

Although the selected structure can serve as a model of the target structure, the orientation and location of the model in the asymmetric unit still need to be found. There are six degrees of freedom to find the correct orientation and location of the

model structure inside the asymmetric unit. Three degrees of freedom correspond to rotation and three degrees of freedom correspond to translation.

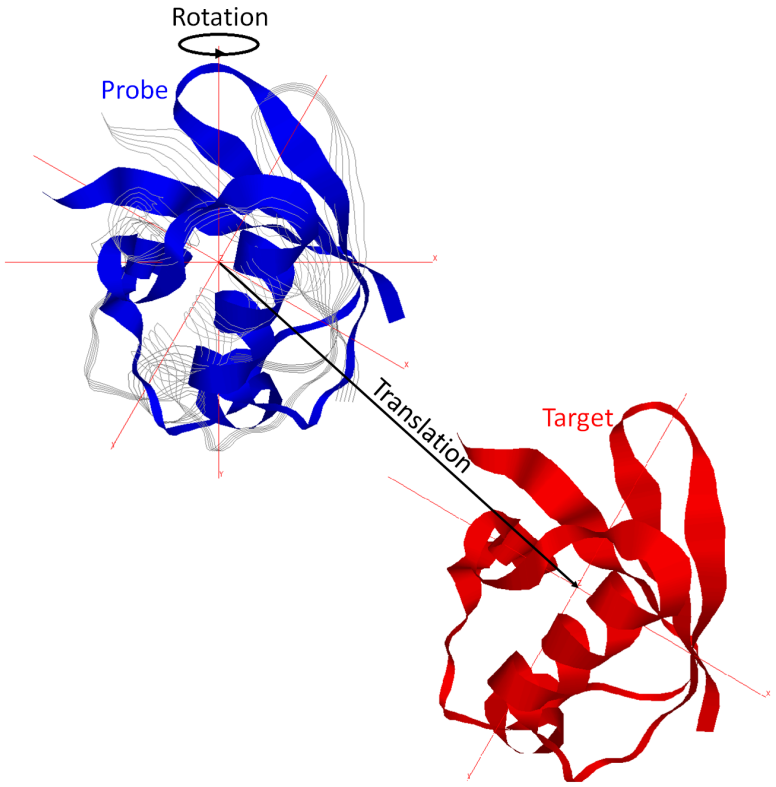

FIGURE 3.12: Illustration of the rotation and translation functions applied to superimpose a probe and target structure in molecular replacement.

$$X' = RX + T \tag{3.9}$$

The orientation and location are searched separately. If the orientation and location are searched at the same time, there are a lot of possible combinations and it requires too much time to compute all possible combinations. Patterson map is used to determine the correct orientation of the model structure in the asymmetric unit. If the orientation of the model structure is correct, there will be maximum overlap between the Patterson

map of the target structure and the Patterson map of the model structure. An overlap function or rotation function is defined as

$$R(\phi, \psi, \chi) = \int_{u,v,w} P_{\text{target}}(u, v, w) P_{\text{model}}(u, v, w, \phi, \psi, \chi) du dv dw \tag{3.10}$$

At each set of rotation angles, the value of the rotation function is an integral of the product of two Patterson functions. When the orientation of the model molecule is correct, the peaks on the two Patterson maps will reach maximum overlap and the rotation function will obtain a maximum value.

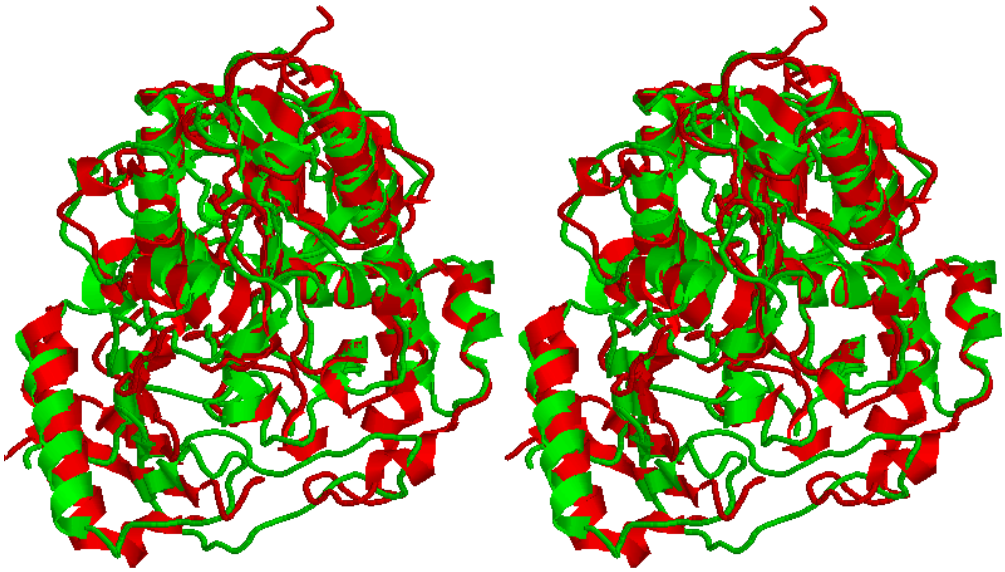

FIGURE 3.13: The model in green and the final structure in red.

After the orientation and the location of the model are determined, the model is placed into the asymmetric unit. Phases are calculated from the model.

The overall agreement between the reflection magnitude of the model and the reflection magnitude of the target can serve as a criterion to identify the correct location

of the model structure in the asymmetric unit. After each translation, the synthetic structure factor of the model is calculated. R factor is defined to calculate the difference between the calculated structure factors of the model and the observed structure factors of the target.

$$R = \frac{\sum \big| |F_{obs}| - |F_{cal}| \big|}{\sum |F_{obs}|} \tag{3.11}$$

For each reflection, the absolute difference between the observed magnitude and the calculated magnitude is calculated. The sum of the absolute difference is calculated over all reflections and then it is divided by the sum of all observed magnitudes. If the location of the model structure is correct, the calculated magnitudes will agree well with the observed magnitudes and R will reach its minimum value.

The standard linear correlation coefficient CC is also used.

$$CC = \frac{\sum_{h,k,l} \left(|F_{obs}|^2 - \overline{|F_{obs}|^2} \right) \times \left(|F_{cal}|^2 - \overline{|F_{cal}|^2} \right)}{\left\{ \sum_{h,k,l} \left(|F_{obs}|^2 - \overline{|F_{obs}|^2} \right)^2 \sum_{h,k,l} \left(|F_{cal}|^2 - \overline{|F_{cal}|^2} \right)^2 \right\}^{1/2}} \tag{3.12}$$

The advantage of this correlation coefficient is that it is scaling insensitive.

3.6 Direct method

Methods which solve the phase problem directly from the observed diffraction intensities are called direct methods[27, 30–49]. The methods generally exploit constraints or statistical correlations between the phases of different reflections. In small molecule crystallography, direct methods are standard techniques for determining the phase angles of the structure factors. The basic assumptions of the direct methods include: the electron density is always positive and the molecular structure consists of discrete atoms.

Phase relations based on probability theory have been formulated and these relations are applied to suitably selected cluster of structure factors. Although direct methods work very well for small molecule crystals, they have not been successfully applied to protein crystals which contain thousands of atoms.

Direct methods are roughly divided into reciprocal-space direct methods and real-space direct methods. Reciprocal-space direct methods focus on the relationships between the phases of different reflections. Real-space direct methods try to use the electron-density constraints.

Reciprocal-space direct methods for small molecule crystals make use of the relationships between phases. Sayre's equation gives us the relationships between the phases of structure factors. Suppose all atoms are the same.

$$\rho(x,y,z) = \frac{1}{V}\sum_{h,k,l} F(h,k,l) exp\big[-2\pi i(hx+ky+lz)\big] \tag{3.13}$$

$$F(h,k,l) = f\sum_{j=1}^{N} exp\big[2\pi i(hx_j+ky_j+lz_j)\big] \tag{3.14}$$

f is the atomic scattering factor which is the same for all atoms. If the electron density function is squared, the new structure factor is denoted as $G(h,k,l)$.

$$\rho^2(x,y,z) = \frac{1}{V}\sum_{h,k,l} G(h,k,l) exp\big[-2\pi i(hx+ky+lz)\big] \tag{3.15}$$

$G(h,k,l)$ is the structure factor of the squared electron density function. It can be proved that $G(h,k,l)$ is the sum of the products of pairs of structure factors whose indices sum

to the desired values of (h, k, l).

$$G(h,k,l) = \frac{1}{V} \sum_{h',k',l'} F(h',k',l') F(h-h', k-k', l-l') \tag{3.16}$$

If we assume that the electron density of neighbor atoms do not overlap with each other, the squared electron density function of the unit cell $\rho^2(x, y, z)$ can be treated as the sum of squared atoms. The structure factor of squared atoms is written as follows.

$$G(h,k,l) = g \sum_{j=1}^{N} exp\big[2\pi i(hx_j + ky_j + lz_j)\big] \tag{3.17}$$

g is the form factor of the squared atom, which is the same for all atoms. So $G(h, k, l)$ can be expressed by $F(h, k, l)$.

$$G(h,k,l) = \frac{g}{f} F(h,k,l) \tag{3.18}$$

Then we got Sayre's equation:

$$F(h,k,l) = \frac{f}{gV} \sum_{h',k',l'} F(h',k',l') F(h-h', k-k', l-l') \tag{3.19}$$

A structure factor is calculated as the sum of the products of pairs of structure factors whose indices sum to the desired values of Miller indices.

$$\begin{aligned} &|F(h,k,l)|exp(i\varphi_{h,k,l}) \\ &= \frac{f}{gV} \sum_{h',k',l'} \big|F(h',k',l')\big|\big|F(h-h', k-k', l-l')\big|exp\big[i\big(\varphi_{h',k',l'} + \varphi_{h-h',k-k',l-l'}\big)\big] \end{aligned} \tag{3.20}$$

Weak reflections contribute a little to the sum. If $|F(h,k,l)|$ is large, $F(h,k,l)$ approximately has the same phase as the terms with large $|F(h',k',l')||F(h-h',k-k',l-l')|$. This method is a powerful way of finding the phases of related strong reflections.

$$\varphi_{h,k,l} = \varphi_{h',k',l'} + \varphi_{h-h',k-k',l-l'} \tag{3.21}$$

It can be written as a triplet relation[50].

$$\varphi_{-h,-k,-l} + \varphi_{h',k',l'} + \varphi_{h-h',k-k',l-l'} \approx 0 \tag{3.22}$$

There is an important assumption in Sayre's equation. The assumption is that the electron densities of neighbor atoms do not overlap. It means atoms should be clearly separated. In practice, the electron densities of neighbor atoms always overlap. But Sayre's equation can still serve as a good approximation[51].

For small molecule crystals, the well-known tangent formula is often used for phase refinement.

$$tan(\varphi) = \frac{\sum_{h',k',l'} \big|E(h',k',l')\big|\big|E(h-h',k-k',l-l')\big| \sin\big(\varphi_{h',k',l'} + \varphi_{h-h',k-k',l-l'}\big)}{\sum_{h',k',l'} \big|E(h',k',l')\big|\big|E(h-h',k-k',l-l')\big| \cos\big(\varphi_{h',k',l'} + \varphi_{h-h',k-k',l-l'}\big)} \tag{3.23}$$

E is the normalized structure factor.

In addition to reciprocal-space direct methods, there are also some famous dual-space direct methods which are used to solve small crystals, such as Shake-and-Bake (SnB) method[52–59] which alternates phase refinement in reciprocal space with density modification in real space, as well as SHELX-D (also called 'Half-Baked' method).

At present, direct methods are used to solve small-molecule crystals. For protein crystals, because of large amount of atoms in the unit cell, in most cases direct methods can't solve the phase problem. However, there are still some real-space direct methods which can be used to locate the protein molecule or even secondary structures in the unit cell.

The real-space approach is a scheme of generating approximate density maps directly from the diffraction intensities[46, 60]. The unit cell is discretized into grid points. On each grid point, one can assign one or zero corresponding to a point scatterer or none. For any given on figuration of the lattice (the pattern of zero's and one's) and a selected diffraction data set, the R value is calculated. By flipping the zero and one into each other on each grid point, one can optimize the R value of the configuration. In essence, this is a construction of optimal binary maps directly from the diffraction intensities. More details can be found in Su (2008), where examples of protein envelopes constructed from real diffraction data can be found.

Initially, some of the grid points are randomly selected to have the value one. The remaining grid points have value zero. The structure factor is calculated and R factor is defined to show the difference between the calculated structure factor and observed structure factor.

$$R = \frac{\sum \left(|F_{obs}| - k|F_{cal}|\right)^2}{\sum |F_{obs}|^2} \tag{3.24}$$

The scale factor k is analytically derived from $\partial R/\partial k = 0$.

$$k = \frac{\sum |F_{obs}||F_{cal}|}{\sum |F_{obs}|^2} \tag{3.25}$$

Initially R is very large. A grid point is randomly selected, if flipping its value between zero and one reduces R value, the value of this grid point will be updated. After thousands of flipping, R is dramatically reduced. When R is reduced, those grid points with value one are capable of showing the location and the envelope of the protein molecule inside the unit cell. Sometimes, the secondary structures can be identified.

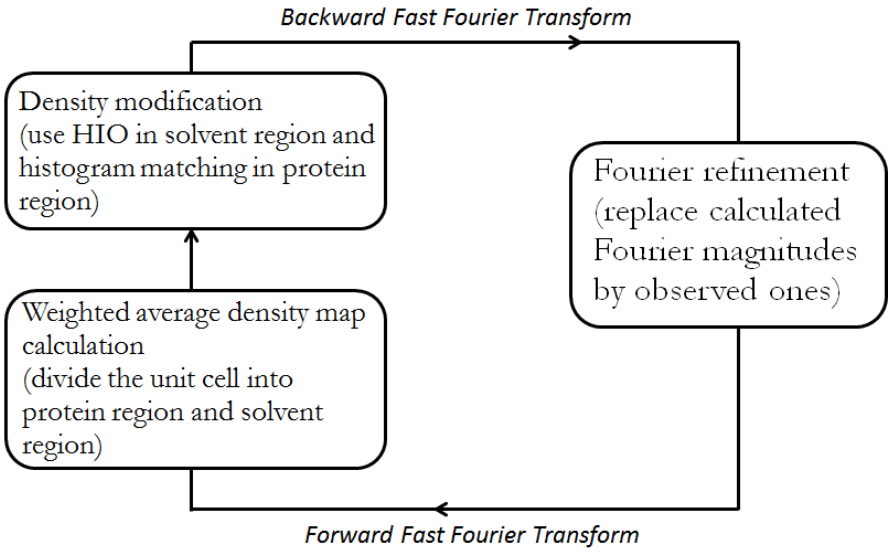

FIGURE 3.14: Simple flowchart of the iterative transform method.

In Chapter 4, a new direct method will be proposed to solve the phase problem for protein crystals. The new method is an iterative transform method. Because the average density in the protein region is higher than the average density in the solvent region, a weighted average density map is used to locate the protein region directly from the observed diffraction intensities. The density in the solvent region is modified by the hybrid input-output algorithm which contains a negative feedback term. The calculated protein-density histogram is modified to match a reference protein-density histogram. After tens of thousands of iterations, the phase problem can be solved for protein crystals with high solvent content directly from the observed intensities. This new direct method

is roughly a real space method at present. Phase improvment techniques such as Sayre's equation will be employed in the future.

3.7 Summary

Some experimental phasing methods have been introduced, including SIR, MIR, SAD and MAD. Experimental phasing methods often need biochemists to prepare several heavy-atom derivative crystals which are time and effort consuming.

Molecular replacement becomes very popular when a similar structure is available to serve as the initial model[61]. If a similar structure is not available, experimental phasing methods are often the only choice.

Direct methods seek to retrieve the phases directly from the observed magnitudes. In this case, biochemists do not need to prepare heavy-atom derivative crystals. They only need to prepare the native protein crystal. The phases are retrieved directly from the diffraction pattern of the native crystal.

In a word, the phase problem can be solved by the application of either isomorphous replacement or molecular replacement or multiple wavelength anomalous dispersion. For small molecular crystallography, direct methods are the standard techniques to determine the phases. However, for protein crystallography, the traditional direct methods have not been successfully applied because of large amounts of atoms in the structure.

In the next chapter, we are going to introduce our new phasing method which is an iterative transform method and it has been successfully tested to solve the phase problem of several protein crystals with high solvent content.

Chapter 4

Iterative transform method

4.1 Introduction

Finding the phases of diffracted X-rays is an important step in protein structure determination. Although the use of selenomethionine and multiple anomalous dispersion (MAD) has rendered the procedure almost routine, the time and resources involved can still be substantial for many large proteins, not to mention the difficulty of expressing some selenomethioninesubstituted proteins in eukaryotic hosts [62]. Alternate techniques that reduce the experimental and investigator demands are therefore still of considerable importance.

Recently, an iterative transform algorithm has been proposed by Liu et al. (2012)[63] to retrieve the phases. An envelope of the region occupied by the protein inside the unit cell is assumed. In each iteration, Fourier refinement (replacing calculated Fourier amplitudes by observed ones) is combined with a density modification in real space, which is essentially a gradual solvent flattening through the hybrid input–output (HIO)

algorithm [64]. For a peculiar choice of the protein boundary, Liu et al. were able to recover several high-resolution structures with high solvent content. As such, that work constitutes important progress in solving the phase problem. It is obviously desirable to eliminate the requirement of a prior knowledge of the protein region. As will be demonstrated below, it is possible to do so and therefore possible to directly phase protein crystals with high solvent content.

We have followed basically Liu et al.'s algorithm except that we allow the protein boundary to evolve with iteration[65]. In each iteration cycle, a weighted average density map is constructed to define the protein region. Thus the protein boundary is not assumed beforehand; rather it is dynamic and becomes accurate only at the end of successful calculations. Therefore, our procedure is ab initio phasing. A very similar idea has been pursued by Millane & Stroud (1997)[66] and by van der Plas & Millane (2000)[67] in their reconstruction of icosahedra virus images from Fourier intensities. Also the related idea of a dynamic support has been studied in the field of coherent diffraction imaging[68, 69].

Although our primary interest is in ab initio phasing, our method can also be used for phase extension. Prior knowledge of low-resolution phases (say 10Å) leads to fast convergence of high-resolution structures.

In this chapter, methodology and techniques are discussed. In Section 2, the over-sampling condition is described. In Section 3, the flow chart of the iterative transform method is given. In Section 4, symmetry operations, equivalent positions and origin choices are discussed. In Section 5, fast Fourier transform method is introduced. In Section 6, weighted average electron-density map is defined to locate the protein mask.

In Section 7, histogram matching is shown as a standard density modification technique. In Section 8, hybrid input-output method (HIO) is employed to modify the electron density in the solvent region. In Section 9, traditional solvent flattening is introduced. In Section 10, strategies to reconstruct missing reflections are discussed. In Section 11, R factors used to monitor the calculated results are given. The last section is a summary of this chapter.

Several examples will be presented in Chapter 5.

4.2 Oversampling condition

In X-ray crystallography, Miao et al.[70] suggest that, given the magnitude of a Fourier transform sampled at the Bragg density, the phase problem is underdetermined by a factor of 2 for three-dimensional crystals. Thus oversampling the magnitude of a Fourier transform by a factor of 2 is required in order to retrieve the phases.

The diffraction pattern with $N_1 \times N_2 \times N_3$ reflections in h, k, and l directions can be observed in reciprocal space. According to Friedel's law, the magnitudes of the structure factors have central symmetry. Therefore, the number of independent reflections drops to $N_1 \times N_2 \times N_3/2$. If the phases of these reflections are solved, the electron density can be calculated on $N_1 \times N_2 \times N_3$ points along the x, y, and z directions in real space. Generally speaking, we have to solve the electron densities on a $N_1 \times N_2 \times N_3$ grid, but we only have $N_1 \times N_2 \times N_3/2$ independent intensities measured in experiment. However, if the solvent content in the crystal is greater than 50%, only less than $N_1 \times N_2 \times N_3/2$ unknown electron densities have to be determined for the protein region. In this case, the number of independent intensities measured in experiment can exceed the number of

unknown electron densities in a unit cell. Therefore, oversampling condition is satisfied and the phase problem becomes overdetermined.

In the iterative transform method, we deal with crystals with solvent content greater than 50% in order to satisfy the oversampling condition.

In Chapter 6, we will discuss that there are many density constraints and phase constraints which can be employed to reduce the degrees of freedom of the electron density. Therefore, low-solvent-content crystals can satisfy the oversampling condition.

4.3 Flowchart of the iterative transform method

An iterative transform method has been proposed to directly solve the phase problem from the X-ray diffraction intensities. This method starts from random density in an asymmetric unit of the unit cell. By backward fast Fourier transform, the calculated phases are combined with observed magnitudes to form new structure factors. The forward fast Fourier transform of these assembled structure factors gives a new electron density map in the unit cell. A weighted average electron-density map is calculated from the new electron density map. A cutoff value can be searched on the weighted average electron density map to divide the unit cell into protein region and solvent region. Inside the protein region, histogram matching is employed to modify the calculated electron density. In the solvent region, hybrid input-output method is used to modify the calculated electron density. After thousands of iterations, the final calculated electron density becomes interpretable for a successful run.

Our iterative transform algorithm is represented by the flowchart in Figure 4.1. The unit cell is divided into a grid with regularly spaced points. At the beginning of the

iterative transform method, the electron density on each grid point in an asymmetry unit is given a random value between zero and one.

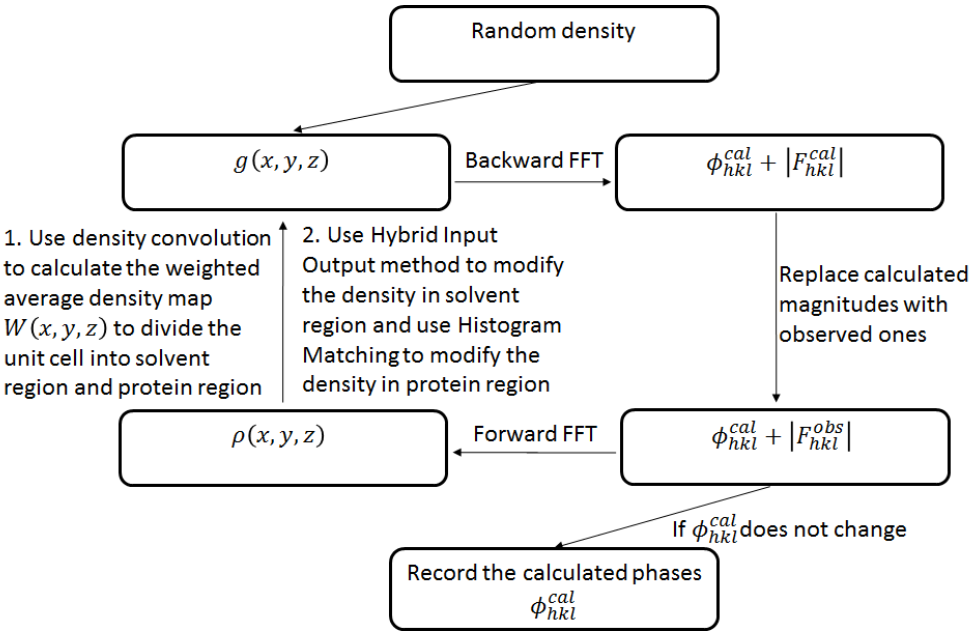

FIGURE 4.1: Flow chart of the iterative transform method.

Each iteration begins with a real space density which comes from the previous round of calculation. By a backward fast Fourier transform of the electron density, the Fourier magnitudes and phases can be calculated. The calculated phases are kept, but the calculated Fourier magnitudes are replaced by the observed Fourier magnitudes. Sometimes this is called Fourier refinement. The calculated phases can also be modified according to constraints in Fourier space such as Sayre's equation. The calculated phases are combined with the observed Fourier magnitudes to produce the new electron density via a forward fast Fourier transform.

Initially, the calculated density in the unit cell is almost random. It is difficult to tell the boundary between the protein and the solvent. However, the average density in the protein region should be a little higher than the average density in the solvent region.

This information is about the overall density distribution in the unit cell and it can be retrieved directly from the observed magnitudes based on my experience. In order to find the boundary between the protein and the solvent, a weighting function is defined. This function does not focus on the density of any local grid point. Actually it focuses on the density distribution in a big region. Therefore, the weighted average density can locate the protein region in the unit cell directly from the observed magnitudes assembled with nearly random phases.

A weighted average density map described above is computed on the calculated electron density map. Because the protein region has higher density than the solvent region, the protein region should have higher weighted average density and the solvent region should have lower weighted average density. A cutoff value of the weighted average density can be found by try and error method to agree with the solvent content of the crystal. The cutoff weighted average density divides the unit cell into two regions. One is the protein region and the other is the solvent region. Both of these two regions are connected regions due to the property of the weighted average density function. The connectivity constraint of the protein region is satisfied. The initial protein boundary is only an estimate. The accuracy of the protein boundary can be realized after thousands of iterations.

There are different density constraints in the protein region and the solvent region. Density constraints are applied via density modification to two regions, respectively.

In the protein region, histogram matching is used to modify the electron density. Although different molecules have different structures, the electron density distributions in the molecular region always look similar at the same resolution. This property can

be used to modify a poor electron density map and push it toward a good electron density map. Histogram matching can be done in 1D which is the density-histogram matching. The gradient or first order derivative of the electron density function can also be matched. This is called gradient-histogram matching. The combination of density-histogram matching and gradient-histogram matching is called 2D histogram matching. Higher order derivatives of the electron density function can also be matched. However, due to the resolution limit and the effect of temperature factor, higher order histogram matching are not used in practice.

In the protein region, if there is non-crystallographic symmetry (NCS), NCS averaging can be a powerful density modification tool. In some crystals, there are several copies of molecules inside an asymmetric unit. Different copies can be related by rotation and translation non-crystallographic symmetry operations. The electron densities for different copies of the molecule in an asymmetric unit can be averaged. The existence of NCS reduces the degrees of freedom of the protein electron density. In other words, it increases the ratio of the solvent volume to the independent protein volume. Oversampling condition can thus be satisfied for low-solvent-content crystals.

In the solvent region, hybrid input-output (HIO) method is used to modify the electron density. Because most water molecules have different orientations in different unit cells, the electron density in the solvent region is flat and equals a constant. If we let F_{000} float, the expected solvent density can be set to zero. In order to push the calculated electron density towards zero in the solvent region, the hybrid input-output method introduces a negative feedback density. This negative feedback density can modify the calculated electron density in the solvent region and push it slowly

towards zero. The speed can be controlled by the feedback parameter. HIO connects the calculated electron density in the current iteration with the calculated electron density in the previous iteration. It is very capable to overcome the stagnation problem.

In addition to hybrid input-output method, solvent flattening is another density modification technique used in the solvent region. Solvent flattening pushes the electron density in the solvent region directly to zero or a constant. There is a disadvantage that the calculated electron density in the solvent region has no time to evolve. Realizing this disadvantage, I suggest a limited HIO method.

4.4 Symmetry operations, equivalent positions, and origin choices

The iterative transform method starts from random density in the asymmetric unit and all density modification techniques are applied inside the asymmetric unit. The electron density in the asymmetric unit can be extended to the whole unit cell according to the symmetry operations of the specific space group.

Fractional coordinates are often used by symmetry operations. There are two kinds of coordinates commonly used in crystallography. Orthogonal coordinates are often used to show the calculated results in Cartesian system, for example, the atomic coordinates in the pdb file. Fractional coordinates are always used in calculations, because fractional coordinates are independent of cell parameters. The transformation between these two coordinates can be realized by two transformation matrices. $M_{orth-to-frac}$ is the matrix

used to go from orthogonal coordinates to fractional coordinates.

$$M_{orth-to-frac} = \begin{bmatrix} 1/a & -\dfrac{\cos\gamma}{a\sin\gamma} & \dfrac{\cos\alpha\cos\gamma-\cos\beta}{a\mu\sin\gamma} \\ 0 & \dfrac{1}{b\sin\gamma} & -\dfrac{\cos\alpha-\cos\beta\cos\gamma}{b\mu\sin\gamma} \\ 0 & 0 & \dfrac{\sin\gamma}{c\mu} \end{bmatrix}$$

where $\mu = \sqrt{1-\cos^2\alpha-\cos^2\beta-\cos^2\gamma+2\cos\alpha\cos\beta\cos\gamma}$. a, b, c, α, β and γ are cell parameters. $M_{frac-to-orth}$ is the matrix used to go from fractional coordinates to orthogonal coordinates.

$$M_{frac-to-orth} = \begin{bmatrix} a & b\cos\gamma & c\cos\beta \\ 0 & b\sin\gamma & \dfrac{c(\cos\alpha-\cos\beta\cos\gamma)}{\sin\gamma} \\ 0 & 0 & \dfrac{c\mu}{\sin\gamma} \end{bmatrix}$$

Different protein crystals may belong to different space groups which have different symmetry operations. Each symmetry operation can be expressed as a combination of a rotation and a translation. The rotation can be mathematically written as a 3×3 matrix while the translation is a three component vector. A complete unit cell is obtained by applying the rotation matrix and the translation matrix onto the fractional coordinates of the protein atoms in an asymmetric unit. The number of symmetry operations determines the number of equivalent positions inside the unit cell.

Let's take the space group $P2_12_12_1$ as an example to show the symmetry operations[71].

$P2_12_12_1$ is the most popular space group in Protein Data Bank (PDB). More than 23% crystals in PDB fall in this space group. There are four symmetry operations which can generate four equivalent positions in the unit cell. Therefore the asymmetric unit only occupies a quarter of the unit cell. For example, the asymmetric unit can be chosen at $0 < x < 0.5$, $0 < y < 0.5$ and $0 < z < 1.0$. The four symmetry operations are expressed as four matrices. Each of them includes a 3×3 rotation matrix and a translation column vector.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 & 0.5 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0.5 \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & -1 & 0.5 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0.5 \\ 0 & -1 & 0 & 0.5 \\ 0 & 0 & -1 & 0 \end{bmatrix}$$

They correspond to four equivalent positions in the unit cell. $(x, y, z); (-x{+}0.5, -y, z{+}0.5); (-x, y + 0.5, -z + 0.5); (x + 0.5, -y + 0.5, -z)$

For example, a human thyroid hormone receptor with PDB ID 3ILZ[72] is in space group $P2_12_12_1$. The unit cell is shown in Figure 4.2.

Now we pick another space group $P4_32_12$ as another example to show the symmetry operations. $P4_32_12$ ranks as the seventh popular space group in Protein Data Bank. There are eight symmetry operations which produce eight equivalent positions in the unit cell. The asymmetric unit only occupies one eighth of the unit cell. The asymmetric unit is chosen at $0 < x < 0.5$, $0 < y < 0.5$ and $0 < z < 0.5$. The eight symmetry

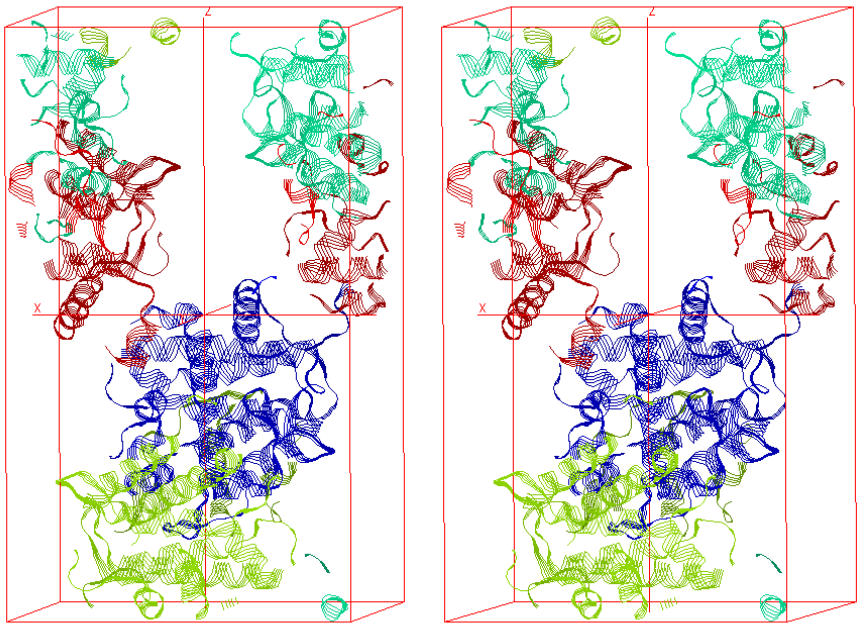

FIGURE 4.2: Stereogram of a unit cell in $P2_12_12_1$ space group, containing four equivalent positions.

operations are expressed as matrices.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0.5 \end{bmatrix} \begin{bmatrix} 0 & -1 & 0 & 0.5 \\ 1 & 0 & 0 & 0.5 \\ 0 & 0 & 1 & 0.75 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0.5 \\ -1 & 0 & 0 & 0.5 \\ 0 & 0 & 1 & 0.25 \end{bmatrix}$$

$$\begin{bmatrix} -1 & 0 & 0 & 0.5 \\ 0 & 1 & 0 & 0.5 \\ 0 & 0 & -1 & 0.75 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0.5 \\ 0 & -1 & 0 & 0.5 \\ 0 & 0 & -1 & 0.25 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0.5 \end{bmatrix}$$

They correspond to eight equivalent positions in the unit cell. $(x, y, z); (-x, -y, z + 0.5); (-y + 0.5, x + 0.5, z + 0.75); (y + 0.5, -x + 0.5, z + 0.25)$ $(-x + 0.5, y + 0.5, -z + 0.75); (x + 0.5, -y + 0.5, -z + 0.25); (y, x, -z); (-y, -x, -z + 0.5)$

As another example, a photosynthetic reaction center structure with PDB ID 2UXJ[73] is crystalized in space group $P4_32_12$. Eight equivalent molecules are generated via symmetry operations to form a complete unit cell, as shown in Figure 4.3.

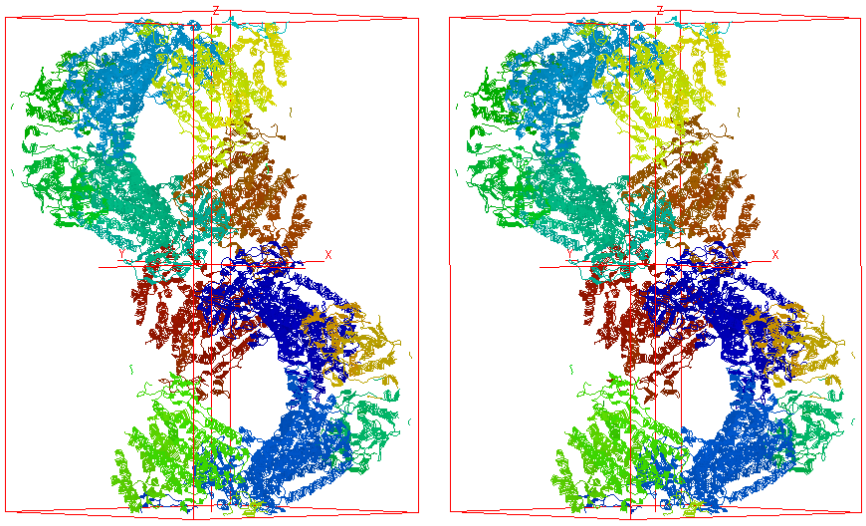

FIGURE 4.3: Stereogram of a unit cell in $P4_32_12$ space group, containing eight equivalent positions.

To apply the iterative transform method, the unit cell has to be divided into a grid. The approximate density for each grid interval should be sampled at the center of that grid interval. In this case, when we apply symmetry operations, the whole unit cell will be properly sampled. For example, we divide the unit cell into a 4×4 grid. Suppose each grid unit is occupied by an atom. The space group is $P2_12_12_1$. In Figure 4.4, the red atoms are located at the center of each grid unit in the asymmetric unit. The green, blue and yellow atoms occupy the equavelant positions in the unit cell. The unit cell is properly sampled after we apply symmetry operations.

However, if the approximate density for each grid interval is sampled at the corner of that grid interval, after applying symmetry operations, the whole unit cell will not

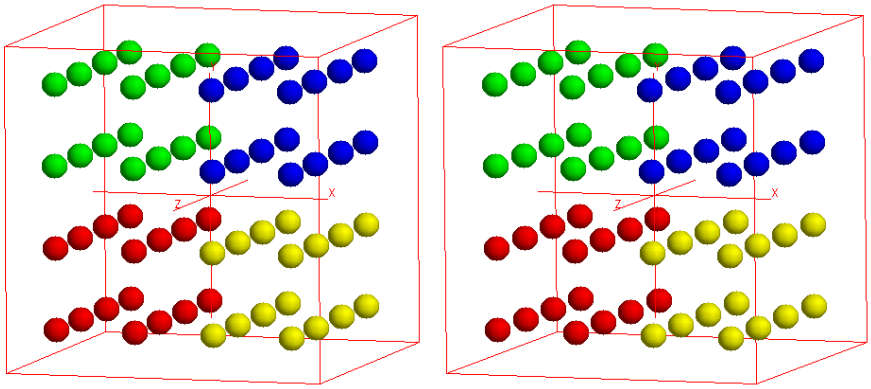

FIGURE 4.4: Stereogram to show the density sampled at the center of each grid interval.

be properly sampled. In Figure 4.5, red atoms are located at a corner of each grid unit in the asymmetric unit. The green, blue and yellow atoms occupy equivalent positions. After symmetry operations, some grid units are doubly sampled, while some grid units are empty. This is obviously not a uniform sampling.

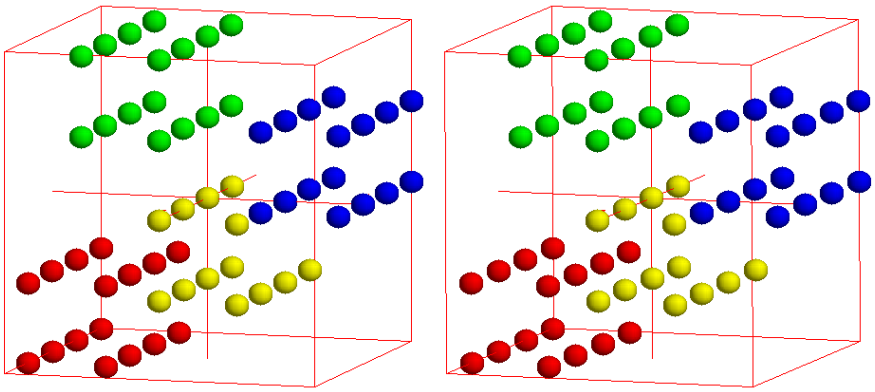

FIGURE 4.5: Stereogram to show the density sampled at the corner of each grid interval.

The unit cell should be divided into a grid and the number of grid units in each

dimension should be divisible by 1, 2, 3, 4, etc., which depends on the symmetry operations of the specific space group. It helps to extend the electron density from an asymmetric unit to the whole unit cell. For example, the unit cell should be divided into even number of grid units in three orthogonal directions if the unit cell has $P2_12_12_1$ symmetry.

Origin choice arises from the fact that the measured intensities are the same for several permissible origin choices of the unit cell[9]. Difference origin choices give different unit cells, but they correspond to the same diffraction pattern. If the protein crystal belongs to a non-centrosymmetric space group, its inverse image gives rise to an identical diffraction pattern. The inverse image is referred to as an enantiomorph. If the enantiomorph exists, the number of origin choices will be doubled. Because all origin choices have the same measured intensities, the calculated electron density can be in any origin choice.

There are sixteen origin choices for crystals in $P2_12_12_1$ space group. The first eight origin choices are (0.5, 0, 0), (0, 0.5, 0), (0, 0, 0.5), (0.5, 0.5, 0), (0.5, 0, 0.5), (0, 0.5, 0.5) and (0.5, 0.5, 0.5). $P2_12_12_1$ space group is non-centrosymmetric. There is no inversion center in the unit cell. Under a coordinate inversion, the magnitudes of the structure factors do not change at all. Therefore, there are another eight origin choices corresponding to enantiomorphs.

There are eight origin choices for crystals in $P4_32_12$ space group. The first four origin choices are (0, 0, 0), (0, 0, 0.5), (0.5, 0.5, 0) and (0.5, 0.5, 0.5). Crystals in $P4_32_12$ space group are non- centrosymmetric. There are another four origin choices due to enantiomorph.

4.5 Fast Fourier Transform

Fast Fourier Transform (FFT) can dramatically reduce the computing time of each iteration cycle. There are several packages that can be used for FFT, such as the Fast Fourier Transform in the West (FFTW) and the Intel Fast Fourier Transform in Intel Math Kernel Library (MKL). Intel MKL FFTs include many optimizations and support a broad variety of FFTs, such as three-dimensional complex-to-complex, real-to-complex and real-to-real transforms of arbitrary length.

Take Intel FFT three-dimensional complex-to-complex transform as an example. The unit cell is divided into a grid with $N_x \times N_y \times N_z$ regularly spaced points. Electron densities on grid points are stored in a $N_x \times N_y \times N_z$ matrix. Because electron densities are real, their imaginary parts should be zero. Structure factors are stored in a $N_h \times N_k \times N_l$ matrix. Structure factors are complex numbers including both magnitudes and phases.

In the manual of Intel FFT, the general form of a d-dimensional discrete Fourier transform is

$$Z_{k_1,k_2,\cdots,k_d} = \sigma \sum_{j_d=0}^{n_d-1} \cdots \sum_{j_2=0}^{n_2-1} \sum_{j_1=0}^{n_1-1} W_{j_1,j_2,\cdots,j_d} exp\left(\delta 2\pi i \sum_{l=1}^{d} j_l k_l / n_l\right) \tag{4.1}$$

for $k_l = 0, \cdots n_l - 1 (l = 1, \cdots, d)$, where σ is an arbitrary real-valued scale factor, and the sign in the exponent is $\delta = -1$ for the forward transform and $\delta = +1$ for the backward transform.

Take a two-dimensional unit cell as an example to demonstrate Intel FFT. The unit cell does not have any symmetry and is divided into a 4×4 grid.

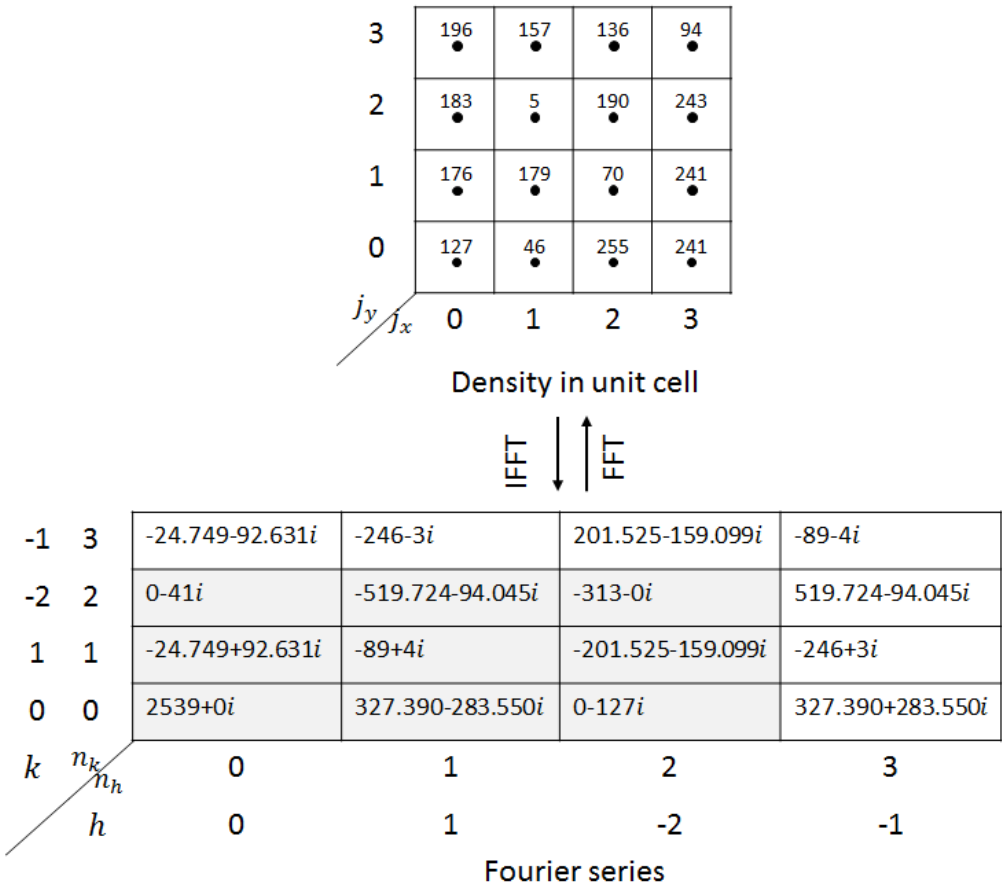

FIGURE 4.6: A two-dimensional example showing the Intel FFT calculation.

According to the previous section, the density for each grid unit is picked at the center of that grid interval. The density of each grid unit has been listed in Figure 4.6. After fast Fourier transform, the calculated Fourier series are stored in a 4×4 matrix. The graph has shown how Fourier series are stored in the matrix. The unique structure factors, which have non-negative indices, only occupy a quarter of the matrix. Other structure factors are derived from the unique structure factors by reciprocal-space symmetry operations and Friedel's law.

Following the general form of the discrete Fourier transform in the manual of Intel FFT, we have to use a backward FFT to transform the electron density in the unit cell

to structure factors. The backward scale factor is $V/(N_xN_yN_z)$.

$$F(h,k,l) = \frac{V}{N_xN_yN_z}\sum_{j_x=0}^{N_x-1}\sum_{j_y=0}^{N_y-1}\sum_{j_z=0}^{N_z-1}\rho(j_x,j_y,j_z)exp\big[2\pi i(hx+ky+lz)\big] \tag{4.2}$$

The calculated magnitudes of the structure factors will be replaced with observed magnitudes. Then we have to use a forward FFT to transform the assembled structure factors to an electron density map in the unit cell. The forward scale factor is $1/V$.

$$\rho(j_x,j_y,j_z) = \frac{1}{V}\sum_{h=-N_h/2}^{N_h/2-1}\sum_{k=-N_k/2}^{N_k/2-1}\sum_{l=-N_l/2}^{N_l/2-1}F(h,k,l)exp\big[-2\pi i(hx+ky+lz)\big] \tag{4.3}$$

$$N_h = N_x; N_k = N_y; N_l = N_z. \tag{4.4}$$

The sizes of those two matrices are the same. The observed data has a resolution. The grid size used to divide the unit cell should be properly chosen so that all observed unique structure factors can be fit into the matrix. Generally, the distance between two nearest grid points is selected from $d_{\text{cutoff}}/4$ to $d_{\text{cutoff}}/2$. Therefore, for orthogonal space groups,

$$\begin{aligned} N_h = N_x &= \frac{a}{d_{\text{cutoff}}/2} \\ N_k = N_y &= \frac{b}{d_{\text{cutoff}}/2} \\ N_l = N_z &= \frac{c}{d_{\text{cutoff}}/2} \end{aligned} \tag{4.5}$$

Each diffraction spot or structure factor has a nominal resolution. The higher the diffraction angle of the spot, the higher is its resolution. In real space, resolution means the distance corresponding to the smallest observable feature. According to Bragg equation

$$d = \frac{\lambda}{2\sin\theta} = \frac{1}{S_{h,k,l}} \tag{4.6}$$

where λ is the X-ray wavelength, and d is the smallest distance between parallel crystal-lattice planes with Miller indices (h, k, l). The resolution is given by d. For a diffraction spot or a structure factor indexed by (h, k, l), its resolution can be calculated as follows.

$$\begin{aligned}
S_{h,k,l} &= \frac{1}{V}\sqrt{h^2S_1 + k^2S_2 + l^2S_3 + 2hkS_4 + 2hlS_5 + 2klS_6} \\
S_1 &= b^2c^2\sin^2\alpha \\
S_2 &= a^2c^2\sin^2\beta \\
S_3 &= a^2b^2\sin^2\gamma \\
S_4 &= abc^2(\cos\alpha\cos\beta - \cos\gamma) \\
S_5 &= ab^2c(\cos\gamma\cos\alpha - \cos\beta) \\
S_6 &= a^2bc(\cos\beta\cos\gamma - \cos\alpha) \\
V &= abc\mu
\end{aligned} \tag{4.7}$$

V is the volume of the unit cell and $\mu = \sqrt{1 - \cos^2\alpha - \cos^2\beta - \cos^2\gamma + 2\cos\alpha\cos\beta\cos\gamma}$

a, b, c, α, β and γ are cell parameters.

After backward FFT, the magnitudes of the calculated structure factors should be replaced by the observed ones. Generally, only unique reflections are recorded in the

experimental data. Hence the assembled structure factors consist of unique structure factors and they have to be extended to the whole reciprocal space by reciprocal space symmetry operations and Friedel's law. The reciprocal space symmetry operations can be derived from the real space symmetry operations. One can also check the tables in the book "International tables for X-ray crystallography Volume I".

4.6 Weighted average density

Inside the unit cell of a protein crystal, there are two regions. One is the protein region occupied by the protein molecule. The other is the solvent region filled by solvent molecules.

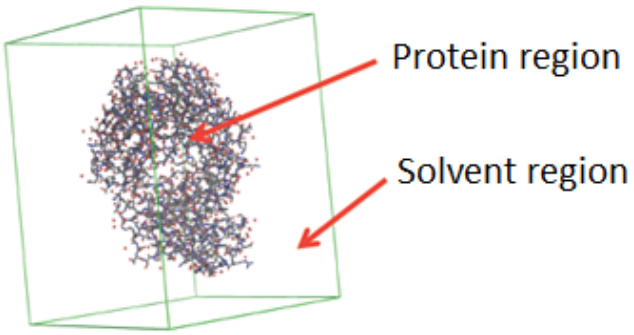

FIGURE 4.7: A unit cell can be divided into protein region and solvent region.

The average density in the protein region should be higher than the average density in the solvent region. Every reflection includes the diffractive contributions from the protein region and the solvent region. In other words, every reflection encodes the information of the protein region and the solvent region and there are tens of thousands of unique reflections. It should be possible to retrieve the protein region from the observed data despite some missing low-resolution reflections. A weighted average density is defined to facilitate this retrieval.

A Gaussian function can be used to serve as a weighting function.

$$w(d_{ij}) = exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right) \tag{4.8}$$

The subscript i or j represents a grid point in the unit cell. Parameter σ measures the width of Gaussian function which can be used to control the convergence of the solvent region. The initial value of σ can be so big that almost all grid points in the unit cell contribute to the average density on a specific grid point. σ is slowly reduced to an appropriate value during the iterations.

We can also use a pyramid function to calculate the weighted average density.

$$w(d_{ij}) = \begin{cases} 1 - d_{ij}/d_0 & d_{ij} \leq d_0; \\ 0 & d_{ij} > d_0. \end{cases} \tag{4.9}$$

d_{ij} is the distance between two grid points. Parameter d_0 characterizes the width of the pyramid function which again can be used to control the convergence of the solvent region. The initial value of d_0 can be as big as half of the unit-cell dimension. During the iterations, d_0 is gradually reduced to an appropriate value.

The weighted average electron density can be calculated in real space.

$$\rho_{ave}(j_x, j_y, j_z) = \sum_{i_x, i_y, i_z} w(d_{ij})\rho(i_x, i_y, i_z) \tag{4.10}$$

Take the pyramid weight as an example. From the weighting function we can see that the closer the grid point is, the bigger weight it has. If a grid point is very far away,

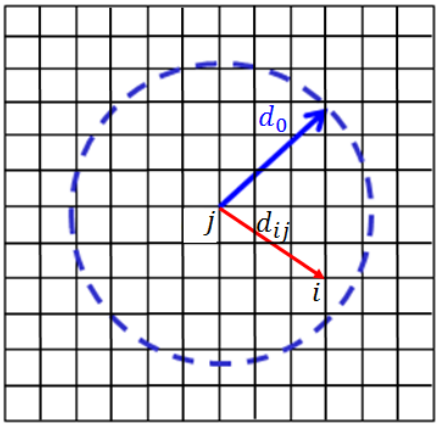

FIGURE 4.8: Schematic diagram to show the calculation of the weighted average density on a grid point in real space.

its weight goes to zero. When we calculate the weighted average density of a grid point, we draw a sphere centered at that grid point. Grid points outside the sphere have zero weight. Different weights are assigned to the grid points inside the sphere. The weights depend on the distance from these grid points to the center of the sphere. If a grid point is closer to the center of the sphere, it has a bigger weight. Otherwise, it has a smaller weight.

If we calculate the weighted average electron density in real space, it will be time consuming. According to the convolution theorem, the weighting can be easily performed in the Fourier space. The Gaussian weighting function can be written as follows.

$$w(r) = exp\left(-\frac{r^2}{2\sigma^2}\right) \tag{4.11}$$

The Fourier transform of the Gaussian weighting function is another Gaussian function.

$$g(S) = (\sqrt{2\pi\sigma^2})^3 exp(-2\pi^2\sigma^2 S^2) \tag{4.12}$$

The pyramid weighting function can be written as

$$w(r) = \begin{cases} 1 - r/d_0 & r \leq d_0; \\ 0 & r > d_0. \end{cases} \tag{4.13}$$

The Fourier transform of the pyramid weighting function can be written as

$$\begin{aligned} g(S) = & \frac{3\big[\sin(2\pi d_0 S) - 2\pi d_0 S \cos(2\pi d_0 S)\big]}{(2\pi d_0 S)^3} \\ & - \frac{3\big\{4\pi d_0 S \sin(2\pi d_0 S) - \big[(2\pi d_0 S)^2 - 2\big]\cos(2\pi d_0 S) - 2\big\}}{(2\pi d_0 S)^4} \end{aligned} \tag{4.14}$$

Upon convolution in Fourier space, each Fourier coefficient gets multiplied by a weight $g(S)$ which only depends on the resolution of the corresponding reflection. After forward fast Fourier transform, the weighted average density on each grid point can be achieved.

$$\rho_{ave}(j_x, j_y, j_z) = \frac{1}{V} \sum_{h=-N_h/2}^{N_h/2-1} \sum_{k=-N_k/2}^{N_k/2-1} \sum_{l=-N_l/2}^{N_l/2-1} g(S) F(h,k,l) exp\big[-2\pi i(hx + ky + lz)\big] \tag{4.15}$$

It is very clear that $g(S)$ can weaken high frequency terms of the Fourier series. In signal processing, they are called Gaussian filter and pyramid filter. They can be classified as low-pass filters, because low frequencies are passed and high frequencies are attenuated. High frequency terms in Fourier series or high resolution structure factors have been attenuated. The pyramid filter attenuates high frequency terms but not as cleanly as what the Gaussian filter does. Basically, Gaussian filter and the pyramid filter

have similar effect.

The protein mask is basically a low-resolution characteristic. Although every reflection memorizes the protein mask in the unit cell, the low-resolution reflections play the most important role. For the synthetic data, we can calculate the exact magnitudes of all reflections. It is very easy and very fast to retrieve an accurate protein mask from the synthetic reflections.

However, some low-resolution reflections are always missing in the experimental data due to the beam stop. Low-resolution reflections have very small diffraction angles. In experiment, the beam stop blocks some low-resolution reflections. We need to rebuild the low-resolution missing reflections in order to locate the protein mask correctly. Different means can be taken to build up the missing reflections. All of them make use of the calculated magnitudes of those missing reflections. For the observed data, because low-resolution diffraction spots have small diffraction angles which can lead to big measurement errors. Those observed low-resolution reflections with big measurement errors should also be rebuilt from the calculated magnitudes.

The protein region inside a unit cell should be connected[74, 75]. However, on the calculated electron density map, those grid points with higher electron density are often separated. The weighted average electron-density map can render the higher-electron-density region connected, as shown in Figure 4.9.

A cutoff value W_{cutoff} can be found by adjusting it such that the calculated solvent content agrees with the expected solvent fraction. Suppose the solvent content is given at the beginning of iterations. The number of grid points in the solvent region can be computed which is the product of solvent content and the total grid points in a unit cell.

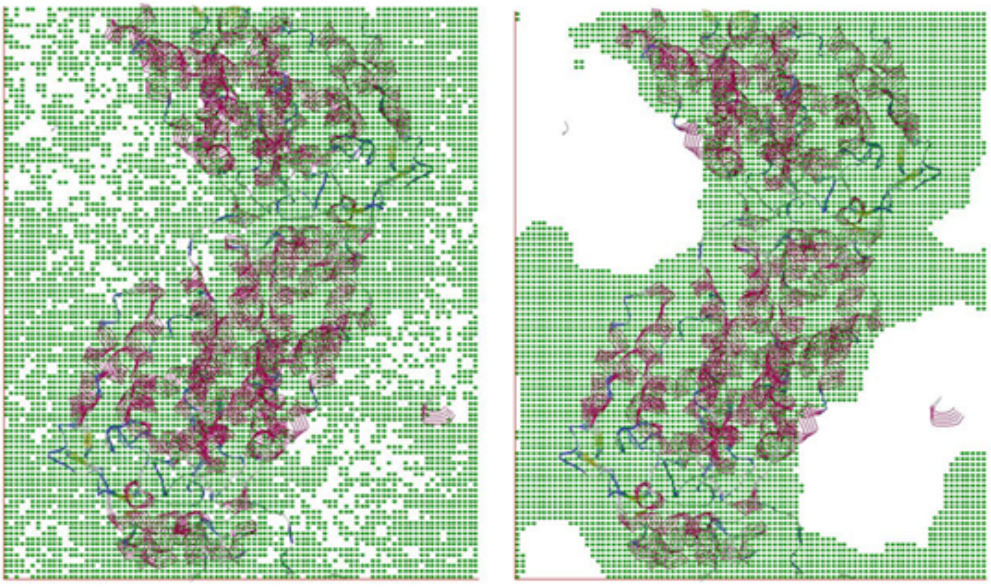

FIGURE 4.9: Electron density map and the corresponding weighted average electron density map. Green regions represent high values. White regions represent small values. The original protein structure has been superimposed in strands.

The cutoff value W_{cutoff} is determined by a half-interval searching algorithm. Let ρ_2 be the maximum weighted average density in the unit cell and let ρ_1 be the minimum weighted average density in the unit cell. Suppose W_{cutoff} equals the average of ρ_2 and ρ_1. The number of grid points with weighted average density greater than W_{cutoff} can be counted. If this number is greater than the number of grid points in the protein region, set ρ_1 equal to the current value of W_{cutoff} and ρ_2 does not change. If this number is less than the number of grid points in the protein region, let ρ_2 be equal to the current value of W_{cutoff} and ρ_1 does not change. This process is repeated until a proper W_{cutoff} is found.

The initial parameter in the weighting function can be as big as half of the unit cell dimension. We can't find an exact protein mask in several iterations. Generally,

hundreds or thousands of iterations are needed and the protein region emerges progressively. The initial calculated protein mask looks like a sphere. The parameter in the weighting function decreases with the iterations. Some details gradually show up on the updated protein mask. After thousands of iteration, the calculated protein mask or protein envelope converges to a fixed profile.

The final value of the parameter in the weighting function should be greater than zero. On the electron density map, there is a small gap between the protein atoms and their neighbor solvent molecules. In this region, the electron density can be less than the average electron density in the solvent region. This region is quite close to the protein surface. They can be treated as part of the protein region. The weighted average density in this region can be less than the one in the solvent region if a small radius in the weighting function is used. The final value of the parameter in the weighting function should be at least several angstroms. If the final value is too big, the calculated protein mask will not be precise enough to give the exact protein shape. If the final value is too small, the protein mask will not be connected and it can't completely cover the protein region. In a word, the final value of the parameter in the weighting function should be properly selected.

4.7 Histogram matching

The frequency distribution of electron density is fairly independent of specific protein structures at the same resolution. The density histogram of a known structure can be used to scale the histogram of a calculated poor density map. This process can improve the calculated density map[76, 77].

The frequency distribution of a reference structure is calculated. The frequency distribution plot is divided into several hundred bins which have equal number of grid points. Suppose the number of bins is N. There are $N+1$ boundaries, starting from 1 to $N+1$. The boundaries of the i^{th} bin is denoted as ρ'_i and ρ'_{i+1} for the reference histogram. Following the same process, ρ_i and ρ_{i+1} are calculated for the poor density map. With a scaling factor a_i, bin width of the poor map can be mapped onto the corresponding bin width of the reference histogram. After a translation, each bin of the poor density map is moved to the correct position on the reference histogram. ρ_i should be moved to ρ'_i , and ρ_{i+1} should be moved to ρ'_{i+1} .

$$\begin{cases} \rho'_i = a_i\rho_i + b_i \\ \rho'_{i+1} = a_i\rho_{i+1} + b_i \end{cases} \tag{4.16}$$

$$a_i = \frac{\rho'_{i+1} - \rho'_i}{\rho_{i+1} - \rho_i} \tag{4.17}$$

$$b_i = \frac{\rho_{i+1}\rho'_i - \rho_i\rho'_{i+1}}{\rho_{i+1} - \rho_i} \tag{4.18}$$

where i ranges from 1 to N. Each bin should have their own a_i and b_i . The calculated electron density on a grid point is located in a bin. The corresponding a_i and b_i of that bin should be used to modify the calculated density.

In practice, the frequency distribution plot is uniformly divided into millions of tiny bins ranging from low electron density to high electron density. In each tiny bin, there are only one or two grid points. It means only one or two grid points are located in the tiny density interval. Starting from the first tiny bin which has minimum density,

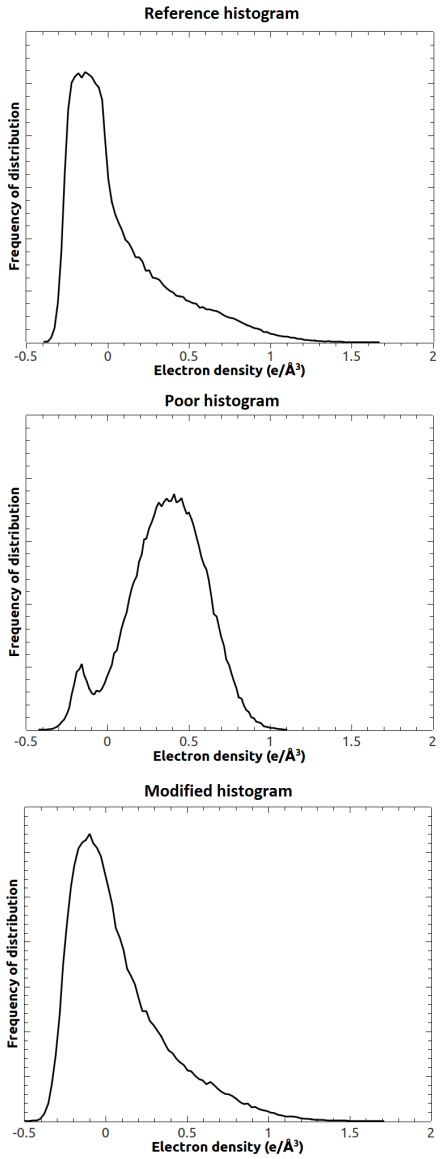

FIGURE 4.10: The reference histogram, the poor histogram and the modified histogram after histogram matching. Histograms are calculated in the protein region at 2Å resolution level.

the number of grid points is counted. The boundaries for each bin as mentioned in the previous paragraph can be easily found. Because the total number of grid points may not be divisible by the total number of bins, the number of grid points in each bin may have a little difference.

Density histogram matching can be applied to the whole unit cell or to the protein

region only. In the iterative transform method, the solvent density is modified by hybrid input-output method which makes the solvent density vibrate around its equilibrium value zero. If density histogram matching is applied in the solvent region, it may be in conflict with HIO. Therefore, histogram matching is only applied to the protein region in the iterative transform method.

4.8 Hybrid input-output

Hybrid input-output (HIO)[64, 78] is a density modification method. Suppose the equilibrium density on a grid point is zero. The calculated density on that grid point often deviates from its equilibrium value. Hybrid input-output method can be employed to drive the calculated density towards its equilibrium value by introducing a negative feedback term. HIO correlates the calculated density in the current iteration with the one obtained in the previous iteration. In protein crystal, the equilibrium density in the solvent region is flat and equals a constant. In the iterative transform method, F_{000} is allowed to change. The equilibrium density in the solvent region can be selected as zero. Therefore, HIO method can be used to modify the calculated density in the solvent region.

In the iterative transform method, suppose $g^{(n)}$ is the electron density at the end of the n^{th} iteration cycle. The $(n+1)^{th}$ iteration cycle starts from a backward fast Fourier transform of $g^{(n)}$. Calculated phases are assembled with observed magnitudes. Then a forward fast Fourier transform is applied onto the assembled structure factors which gives a new electron density $\rho^{(n+1)}$. A weighted average electron-density map is calculated from $\rho^{(n+1)}$ and a cutoff value on the weighted average electron-density map is

found to divide the unit cell into protein region and solvent region. In the protein region, hybrid input-output keeps the density $\rho^{(n+1)}$. In the solvent region, hybrid input-output method introduces a negative feedback density.

$$g^{(n+1)} = \begin{cases} \rho^{(n+1)} & \text{in protein region;} \\ g^{(n)} - \varepsilon\rho^{(n+1)} & \text{in solvent region.} \end{cases} \tag{4.19}$$

ε is a feedback parameter which can be used to optimize the convergence of the algorithm. Empirically, ε is chosen to be around 0.9.

Hybrid input-output method can cause a significant change the calculated electron density in the solvent region. Sometimes, this is not good. We want to control the modified density in the solvent region. A limited hybrid input-output scheme is proposed as follows.

$$g^{(n+1)} = \begin{cases} \rho^{(n+1)} & \text{in protein region} \\ g^{(n)} - \varepsilon\rho^{(n+1)} & \text{if } |g^{(n)} - \varepsilon\rho^{(n+1)}| < \rho_0 \text{ in solvent region} \\ \rho_0 & \text{if } g^{(n)} - \varepsilon\rho^{(n+1)} > \rho_0 \text{ in solvent region} \\ -\rho_0 & \text{if } g^{(n)} - \varepsilon\rho^{(n+1)} < -\rho_0 \text{ in solvent region} \end{cases} \tag{4.20}$$

ρ_0 is a positive parameter which can be used to control the modified density in the solvent region. ρ_0 can be updated during the iterations. For example, the initial value can be set to 1.0 e/Å3 and the final value can be set to 0.2 e/Å3 .

4.9 Solvent flattening

Solvent flattening is a traditional density modification technique[79–81]. It is based on the fact that the electron density in the solvent region should be a constant. If the solvent region can be identified in the asymmetric unit, the calculated electron density in the solvent region can be directly set to zero. Let F_{000} be free to vary. The solvent flattening scheme can then be expressed as follows.

$$g^{(n+1)} = \begin{cases} \rho^{(n+1)} & \text{in protein region} \\ 0 & \text{in solvent region} \end{cases} \tag{4.21}$$

Solvent flattening pushes the solvent density directly to zero. In practice, the experimental data contains noise, temperature factors and measurement errors. The real electron density in the solvent region may not be exactly constant. Based on my experience, the limited hybrid input-output method has better performance than simple solvent flattening when the parameter ρ_0 is given a small value, for example 0.2 e/Å3 .

4.10 Missing reflections

In X-ray diffraction experiment, a beam stop is always needed to stop the intense incident beam that has not been diffracted by the crystal. Otherwise, the detector might be damaged. Those missing reflections have very small diffraction angles and they should be around the center of the diffraction pattern shown in Figure 4.11.

The missing low-resolution reflections are very important for the determination of the protein mask and they have to be rebuilt during the calculations. There are several

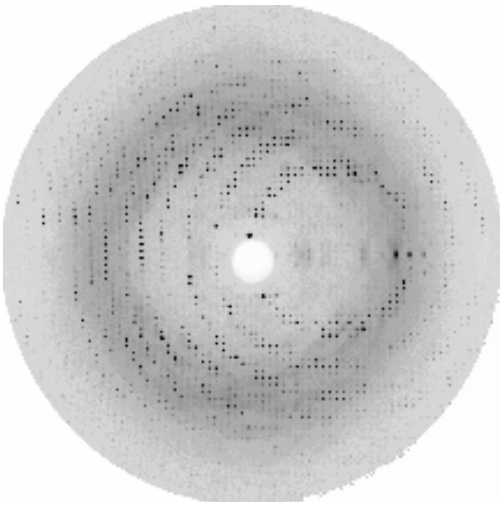

FIGURE 4.11: Some low-resolution diffraction spots on the diffraction pattern are missing due to the beam stop.

means to build up the missing reflections. All of them make use of the calculated structure factors. Suppose regions E represent the observed data. E' stands for the missing regions including the near-forward missing reflections due to the beam top. The missing reflections can be reconstructed from their calculated value by multiplying by a scale factor which is the sum of $|F_{obs}(h,k,l)|$ divided by the sum of $|F_{cal}(h,k,l)|$ in the regions E[63].

$$|F_{obs}(h,k,l)|_{hkl\in E'} = |F_{cal}(h,k,l)|_{hkl\in E'} \frac{\sum_{hkl\in E} |F_{obs}(h,k,l)|}{\sum_{hkl\in E} |F_{cal}(h,k,l)|} \tag{4.22}$$

In addition to the previous scale factor, some other scale factors may be used.

$$|F_{obs}(h,k,l)|_{hkl\in E'} = \xi |F_{cal}(h,k,l)|_{hkl\in E'} \tag{4.23}$$

Sometimes, for some crystals, the scale factor is not needed.

$$|F_{obs}(h,k,l)|_{hkl\in E'} = |F_{cal}(h,k,l)|_{hkl\in E'} \tag{4.24}$$

In phase improvement and model refinement, usually 5% observed data are used to calculate the free R factor. These reflections should also be rebuilt from their calculated structure factors.

4.11 R factor

The traditional measure of the accuracy of the model, is the R factor, sometimes called residual factor or reliability factor. During phase improvement, R factor is used to measure the agreement between the calculated data and the observed data. It is possible to overfit or misfit the diffraction data. In model refinement, an incorrect model sometimes can be refined to a fairly good R value. In order to solve the over-fitting problem, R_{free} has been introduced[5].

At the beginning of phase retrieval, about 5% observed reflections are randomly selected as free data set and they will not be used for phase retrieval. All other observed reflections are in work data set. R_{free} is calculated from the free data set. R_{work} is calculated from the working data set. If the calculated electron-density map is correct, it should predict all the observed data with uniform accuracy. Both R_{work} and R_{free} decrease to a value close to zero. When over-fitting happens, R_{work} decreases to a small value, but R_{free} is still big. Therefore, R_{work} and R_{free} can help us identify an over-fitting density map. A correct density map should have good R_{work} and R_{free} at the same time.

$$R_{\text{free}} = \frac{\sum_{hkl \in \text{free}} \big(|F_{obs}| - k|F_{cal}| \big)}{\sum |F_{obs}|} \tag{4.25}$$

$$R_{\text{work}} = \frac{\sum_{hkl \in \text{work}} \big(|F_{obs}| - k|F_{cal}| \big)}{\sum |F_{obs}|} \tag{4.26}$$

For a random electron-density map, both R_{work} and R_{free} are close to 0.59. For a correct electron density map, R_{free} and R_{work} should be less than 0.30 depending on the resolution of the observed data.

4.12 Summary

In this chapter, an iterative transform method has been introduced. It is a direct phasing method. No prior knowledge of the protein is required. In principle, when the oversampling condition is satisfied, the phase problem can be solved directly from the diffraction pattern of the native protein crystal.

It is crucial to retrieve an accurate protein mask progressively from the diffraction intensities. Weighted average density is employed to locate the protein mask. The initial protein mask is almost a random mask. In each iteration cycle, the protein mask is updated. After thousands of iterations, the final protein mask can be very accurate.

Some techniques used in the iterative transform method have been described, such as fast Fourier transform, histogram matching, hybrid input-output and solvent flattening. Fast Fourier transform makes it possible to complete tens of thousands of iterations in several hours for a large protein molecule. Histogram matching reduces the freedom of the electron density in the protein region. Hence oversampling condition can be better

satisfied. Hybrid input-output method has been proved to be a very effective density modification technique. By introducing a negative feedback parameter, it correlates the calculated density of current iteration with the one of previous iteration. Therefore, it is very capable of overcoming the stagnation problem. Solvent flattening is a traditional density modification technique. It pushes the density in the solvent region directly to zero[82].

Experimental data contains measurement errors and missing central reflections due to the beam stop. Those reflections with big measurement errors and the missing reflections should be reconstructed from the calculated values. Several means have been given to rebuild these reflections.

Limited hybrid input-output method is a new density modification technique which takes the advantages of both HIO and solvent flattening. The density modification in the solvent region by HIO is limited by an adjustable parameter. It is not surprising that the limited HIO has better performance than both HIO and solvent flattening. In Chapters 5 and 6, we will show that the limited HIO method is a better choice for protein crystals with a solvent content close to 50%. The iterative transform method has been successfully applied to solve the phase problem for five protein crystals with high solvent content. Examples will be given in Chapters 5 and 6.

Chapter 5

Direct phasing of protein crystals with high solvent content

5.1 Introduction

In this chapter, the results described in Sections 5.2 and 5.3 have been published[1].

In Chapter 4 we have introduced our iterative transform method. The method starts from a random density map. Each iteration cycle begins with a real-space density map from the previous round of calculation. A backward fast Fourier transform of the density yields phases which are combined with the observed Fourier magnitudes to produce (via a forward fast Fourier transform) the new electron density. A weighted average density map can be derived from the electron density map and a proper cutoff value can be found by adjusting it such that the calculated solvent content agrees with

[1] Hongxing He and Wu-Pei Su, Acta Crystallographica Section A: Foundations and Advances, 71(1):92–98, 2015[65]. According to the copyright of that Journal, the article can be re-used in my thesis as long as a full reference to the paper is given.

the expected solvent fraction. The unit cell has been divided into protein region and solvent region. Histogram matching and hybrid input-output are employed to modify the calculated electron density in the protein region and the solvent region, respectively. After density modification, the next round of iteration begins with the modified electron density. During the iterations, the estimated protein mask is dynamically updated and gradually converges to the correct shape. The correct electron density inside the protein region emerges progressively.

In this chapter, we focus on trial calculations carried out for five structures on their experimental observed data. The crystals of these structures have solvent content greater than 50% where oversampling condition is satisfied. Our method can be easily applied to different space groups. Crystals in two popular space groups $P4_32_12$ and $P2_12_12_1$ are tested to give five examples.

In space group $P4_32_12$, three structures are tried. One is a photosynthetic reaction center structure with PDB ID 2UXJ[73]. The solvent content of this crystal is 76.56%. The resolution of the observed data is 2.25Å. The second structure is a formaldehyde-activating enzyme (Fae) structure with PDB ID 1Y5Y[14]. The solvent content of this crystal is 68.0%. The resolution of the observed data is 2.0Å. The third structure is a flavor protein WrbA from Escherichia coli, with PDB ID 3B6I[83]. The solvent fraction of this crystal is 73.71% and the resolution of the observed data is 1.66 Å.

In space group $P2_12_12_1$, two structures are tried. One is a human thyroid hormone receptor with PDB ID 3ILZ[72]. The solvent content of this crystal is 69.43%. The resolution of the observed data is 2.25Å. The other is a pig pancreatic alpha-amylase with PDB ID 1WO2[84]. The solvent content of this crystal is 70.00%. The crystal

diffracts to 2.01Å.

This chapter is organized as follows. From Section 2 to Section 6, the calculated results of 2UXJ, 1Y5Y, 3ILZ, 1WO2 and 3B6I are presented, respectively. Discussions and conclusions are given in Sections 7 and 8, respectively.

5.2 A photosynthetic reaction center structure with PDB ID 2UXJ

The first structure is a photosynthetic reaction center structure with PDB ID 2UXJ[73]. The space group is $P4_32_12$. The cell dimensions are a = 139.376Å , b = 139.376Å and c = 235.041Å. There are 848 amino acids in the asymmetric unit. The number of non-hydrogen atoms in the asymmetric unit is 7707, including 6487 protein atoms, 817 heterogen atoms and 403 fixed solvent atoms. The solvent content is 76.56%. The crystal diffracts to 2.25Å, with lowest resolution at 27.12Å. There are 103, 927 observed unique reflections. The completeness of the observed data is 94%, with an overall R value of 0.195. It is a good data set, but just like any typical set, there are reflections missing including 92 reflections below 27.12Å. The magnitudes of missing reflections below 2.25Å including F_{000} are automatically reconstructed during the iterations. In other words, the intensities of missing reflections are calculated from the estimated density function in each iteration. The unit cell is discretized into a $140 \times 140 \times 236$ grid for fast Fourier transform. The distance between two nearby grid points is 1Å. The density function is defined only on the grid points.

In principle, we can choose a tight mask for the protein region (i.e. 23% of the unit-cell volume) during each iteration step. But we have found that a somewhat different choice seems to be more effective. We have chosen a loose mask which includes 31% of the unit-cell volume, i.e. the volume of the protein plus 8% solvent. Correspondingly, the solvent region computed from the average density map occupies 69% of the unit-cell volume. This choice is motivated by the thinking that during the iterations, the computed boundary might not match the surface of the protein tightly.

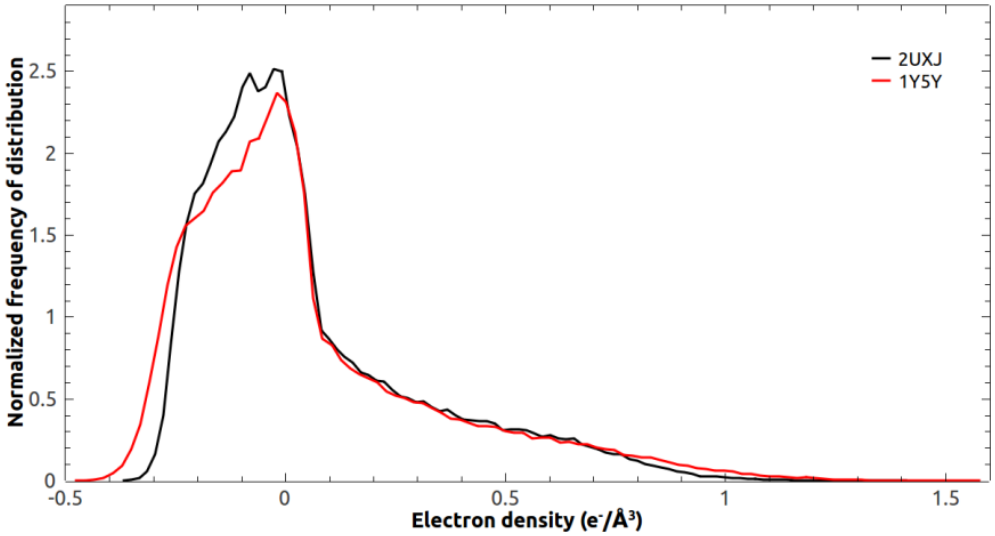

FIGURE 5.1: Density histogram inside a loose protein mask for 2UXJ (black) and 1Y5Y (red) at 2.25Å resolution.

With the choice of a loose protein boundary, the density histogram inside the protein mask is shown as the black curve in Figure 5.1. The corresponding density histogram of a somewhat smaller formaldehyde-activating enzyme (Fae) structure with PDB ID 1Y5Y[14], in the same space group, with a smaller solvent fraction 68%, is shown as the red curve in the same figure. There is a substantial difference between the two histograms. It turns out that they lead to the same result. As a note, the density histograms are calculated in the standard way. In the calculation of the reference density

histogram, a choice of F_{000} is made so that the average density inside the protein mask is about 0.05 e/Å3. This choice was found to work empirically. It partially reflects the solvent contained within the protein mask.

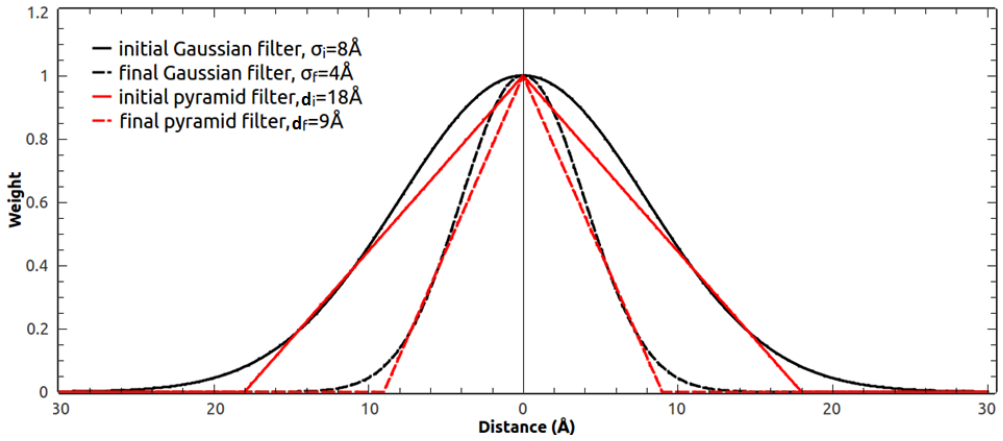

FIGURE 5.2: Weighting functions used for calculating the weighted average density at the beginning and end of an iterative run.

In Chapter 4, we have discussed several weighting functions. After some trial and error, a good choice of the parameter σ of the Gaussian weighting function is found to decrease linearly from 8Å to 4Å in 10,000 iterations. Alternatively, a pyramidal weighting function can be used for which the parameter d_0 decreases from 18Å to 9Å in the same number of iterations, as shown in Figure 5.2. The initial and final values of these parameters can be different for different crystals. Basically, it depends on the cell dimensions and the completeness of the low-resolution data.

The feedback parameter in the hybrid input-output method is taken to be $\varepsilon = 0.9$. This parameter is also used during the phase retrieval of several other crystals which will be described in the following sections. 5% observed reflections are set aside to calculate the free R factor.

To monitor the evolution of the iteration, we compute the mean error in phase angle, defined as follows.

$$\Delta\varphi = \frac{\sum_{h,k,l} \arccos\{\cos[\varphi_{\text{true}}(h,k,l) - \varphi_{cal}(h,k,l)]\}}{\sum_{h,k,l} 1} \tag{5.1}$$

With the above choice of parameters, a batch of 20 independent calculations, with different random starting phases, are carried out. The phase error of the eight successful runs and one unsuccessful run is depicted in Figure 5.3(a). In all the successful runs, the error drops suddenly from 90° to about 50°. After 8000 iterations, the hybrid input-output scheme is gradually turned off and a complete solvent flattening is imposed after 9500 iterations. That leads to a further drop of the phase error to about 32° , whereas for the unsuccessful run, the phase error remains at 90° which are almost random phases. There are also some half-successful runs which have not been shown on the figure. For the half-successful runs, the mean error in phase angles for low-resolution reflections drops a little bit. It implies the calculated protein boundary is somewhat correct.

It is also very instructive to examine the evolution of the R value, which is calculated after density modification in each iteration cycle. As is well known, R_{free} is the R value that correlates well with the phase error[5]. For the R_{free} calculation, 5% of the diffraction intensity data is set aside from the working set. If the completeness of the observed data is not high, fewer reflections should be used as free set, for example 2% reflections. For the same runs depicted in Figure 5.3(a), their R_{free} values are shown in Figure 5.3(b). Clearly, the R_{free} value tracks the phase error in the sudden drop. After solvent flattening, the successful runs all end up with a unique R_{free} value of about 0.23.

The R_{work} value is actually quite informative too, as shown in Figure 5.3(c). When a loose protein mask is used, a sudden drop of R_{work} is observed on a successful run. However, when a tight protein mask is used, there is no sudden drop on R_{work} , but a sudden drop on R_{free} still shows up when a successful run has been reached. Turning off HIO always makes R_{work} decrease. For a successful run, after turning off HIO, R_{work} is still smaller than the value of the failed runs.

Since the sudden drop of the R value is a good indicator of a corresponding improvement in phase error, it is very useful for a new structure determination.

A phase error of 32° means the final density map is very accurate. Examples of 2.25Å maps are shown in Figure 5.4. Some of the water molecules are visible in the map.

It is also of interest to have a look at the protein mask near the end of a successful run (Figures. 5.5 and 5.6). There are still small parts of the protein sticking out of the mask despite the accuracy of the calculated phases.

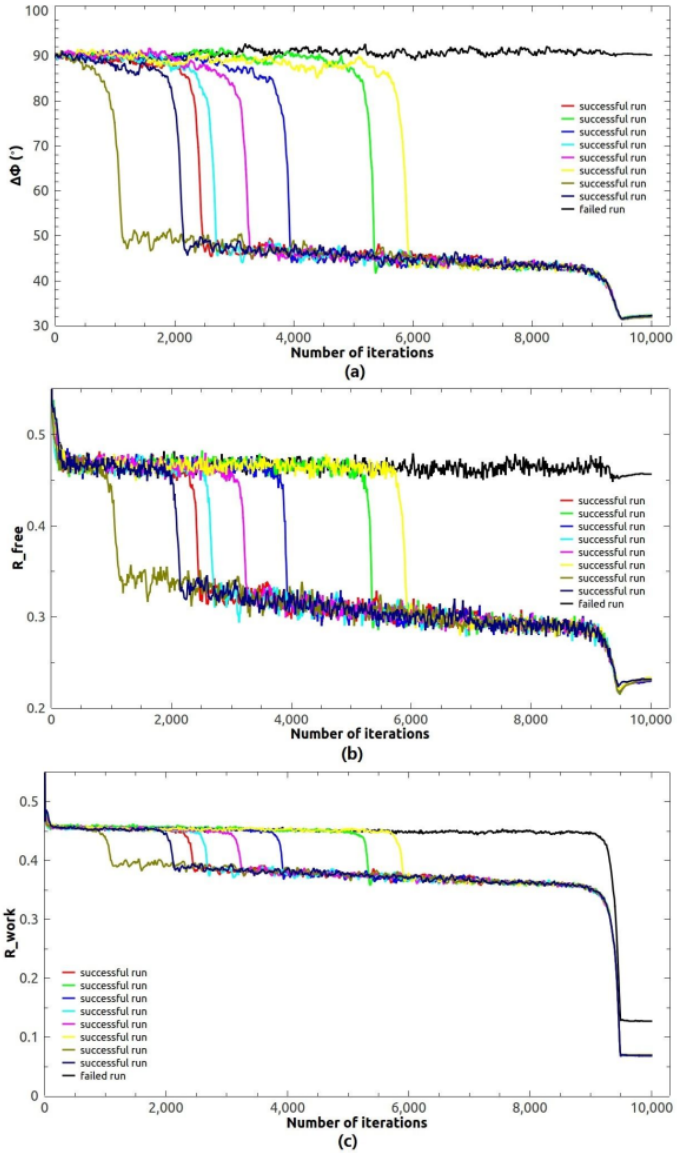

FIGURE 5.3: Evolution of the phase error and R values for eight successful and one unsuccessful runs of 2UXJ.

As an example of our ability to invert lower-resolution data, we cutoff the diffraction data of 2UXJ at 3.5Å resolution and used them in a calculation. A loose protein mask which occupies 31% of the unit-cell volume was updated in iterations. The reference density histogram was computed from 1Y5Y at 3.5Å resolution. The evolutions of the phase error and R_{free} are shown in Figure 5.7. Among a batch of 20 runs, five are

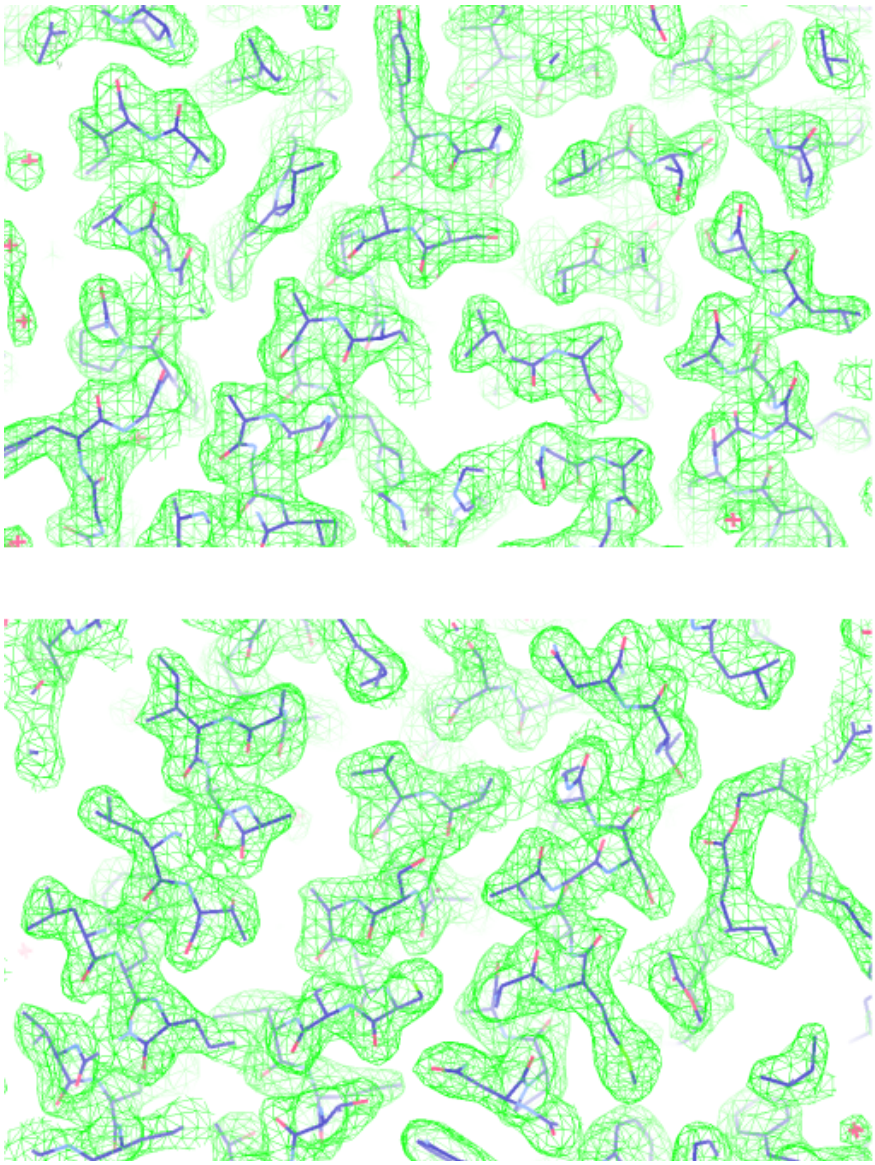

FIGURE 5.4: Two calculated 2.25Å electron-density maps (green) of 2UXJ from successful runs in Figure 5.3. The protein structure of 2UXJ has been superimposed (blue and red).

successful with a mean phase error of 34°. Typical electron-density maps are shown in Figure 5.8.

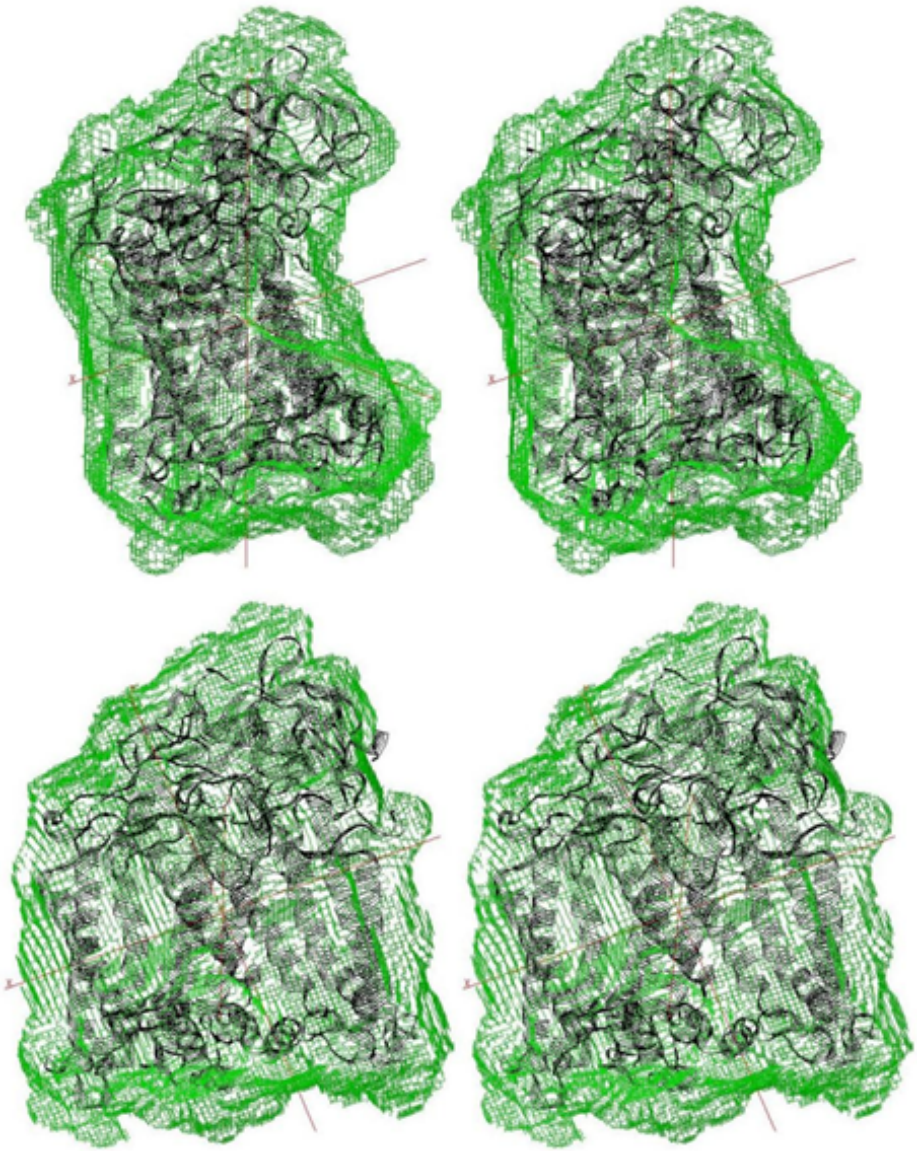

FIGURE 5.5: Stereograms of the calculated protein mask (green) from a successful run in Figure 5.3, compared with the atomic model of 2UXJ in strands (black).

5.3 A formaldehyde-activating enzyme (Fae) structure with PDB ID 1Y5Y

The second structure is a formaldehyde-activating enzyme (Fae) structure with PDB ID 1Y5Y[14]. The space group is $P4_32_12$. The cell dimensions are a = 120.659Å ,

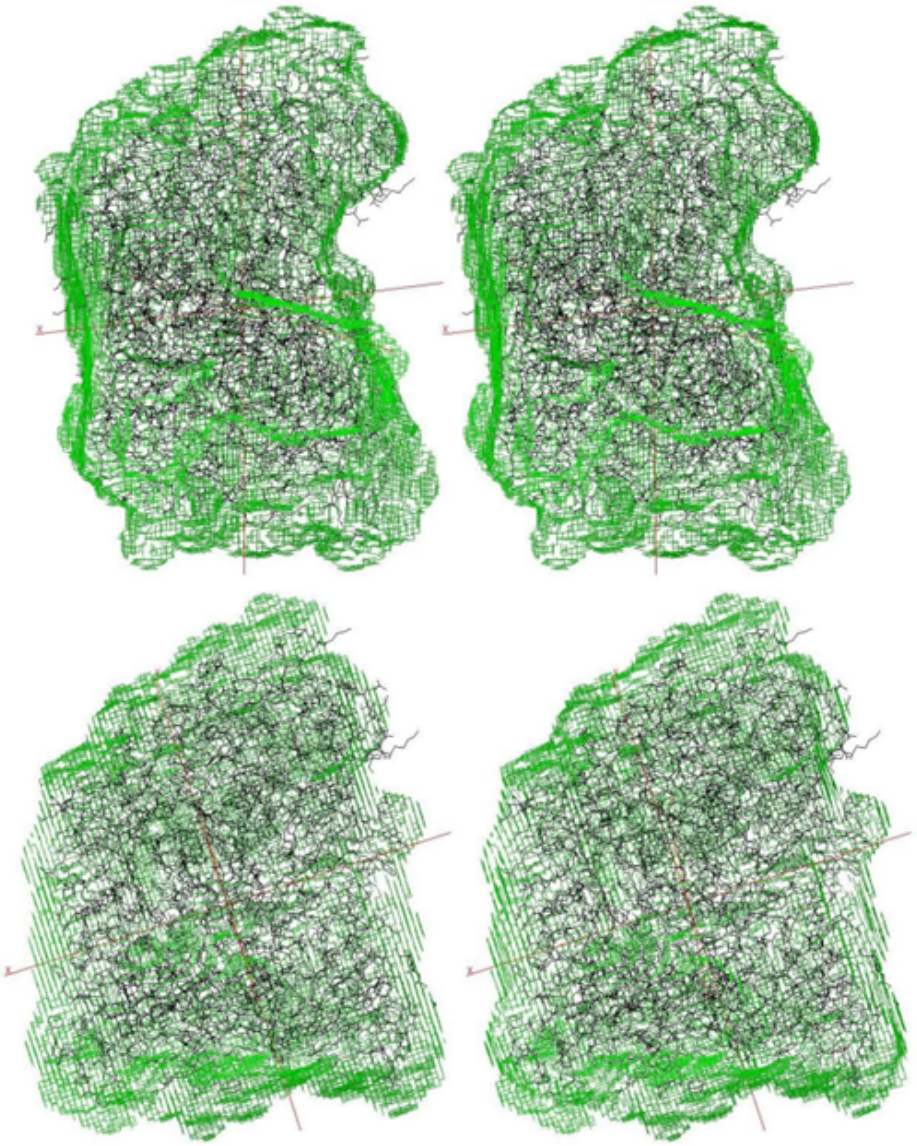

FIGURE 5.6: Stereograms of the calculated protein mask (green) from a successful run in Figure 5.3, compared with the atomic model of 2UXJ in wireframe (black).

b = 120.659Å and c = 205.947Å. The asymmetric unit has C5 non-crystallographic symmetry. There are 845 amino acids in the asymmetric unit. The number of non-hydrogen atoms in the asymmetric unit is 6864. The solvent content is 68.00%. The crystal diffracts to 2.00Å , with lowest resolution at 19.89Å. The number of observed unique reflections is 100,903. The completeness of the data set is 98%, with an overall R

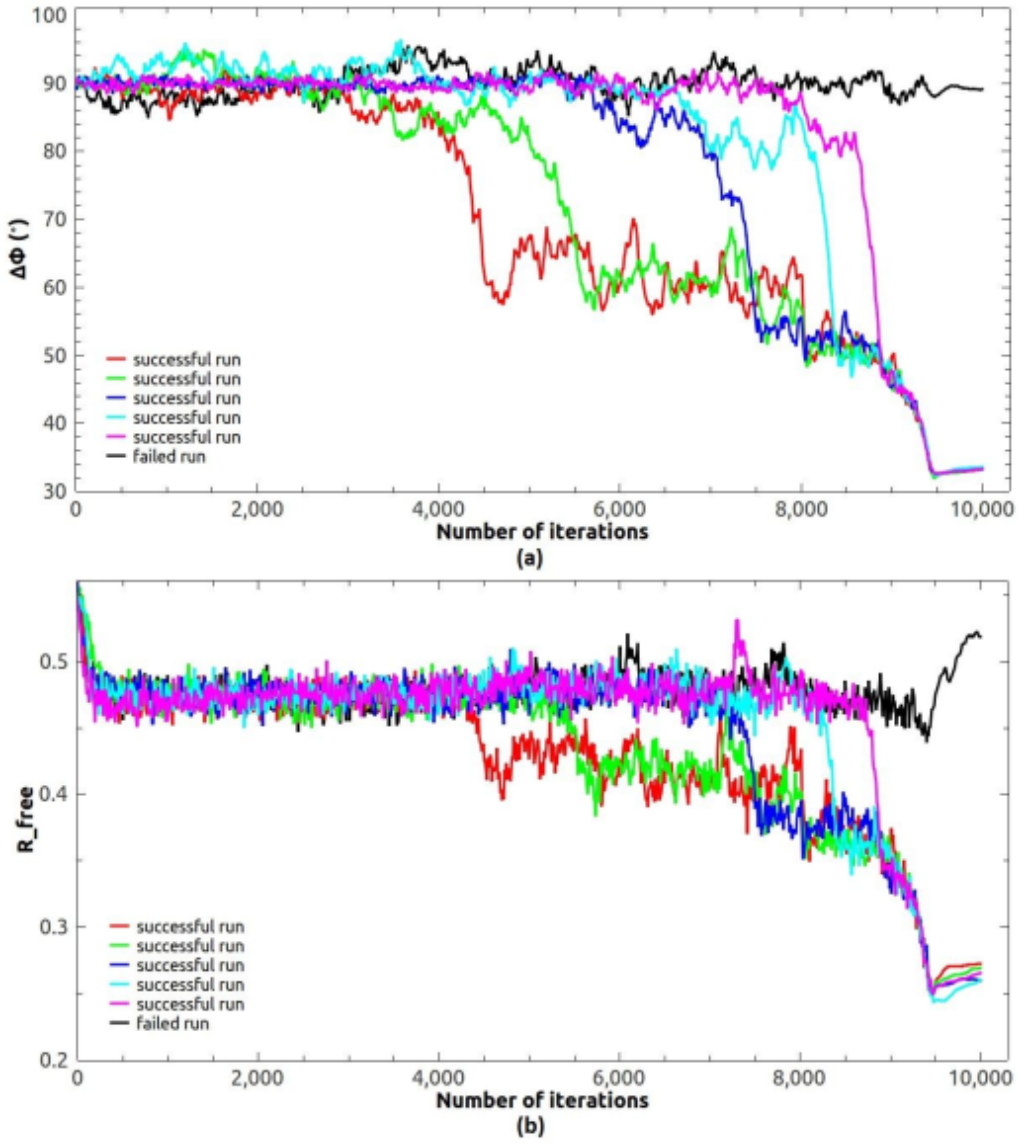

FIGURE 5.7: Evolution of average phase error and R value of 3.5Å calculations of 2UXJ.

value of 0.221. There are about 150 reflections below 19.89Å are missing due to the beam stop. The magnitudes of missing reflections below 2.25Å including F_{000} are automatically reconstructed during the iterations. The unit cell is discretized into a $120 \times 120 \times 208$ grid for fast Fourier transform. The distance between two nearby grid points is 1Å. The density function is defined only on the grid points. In each iteration cycle, a loose mask

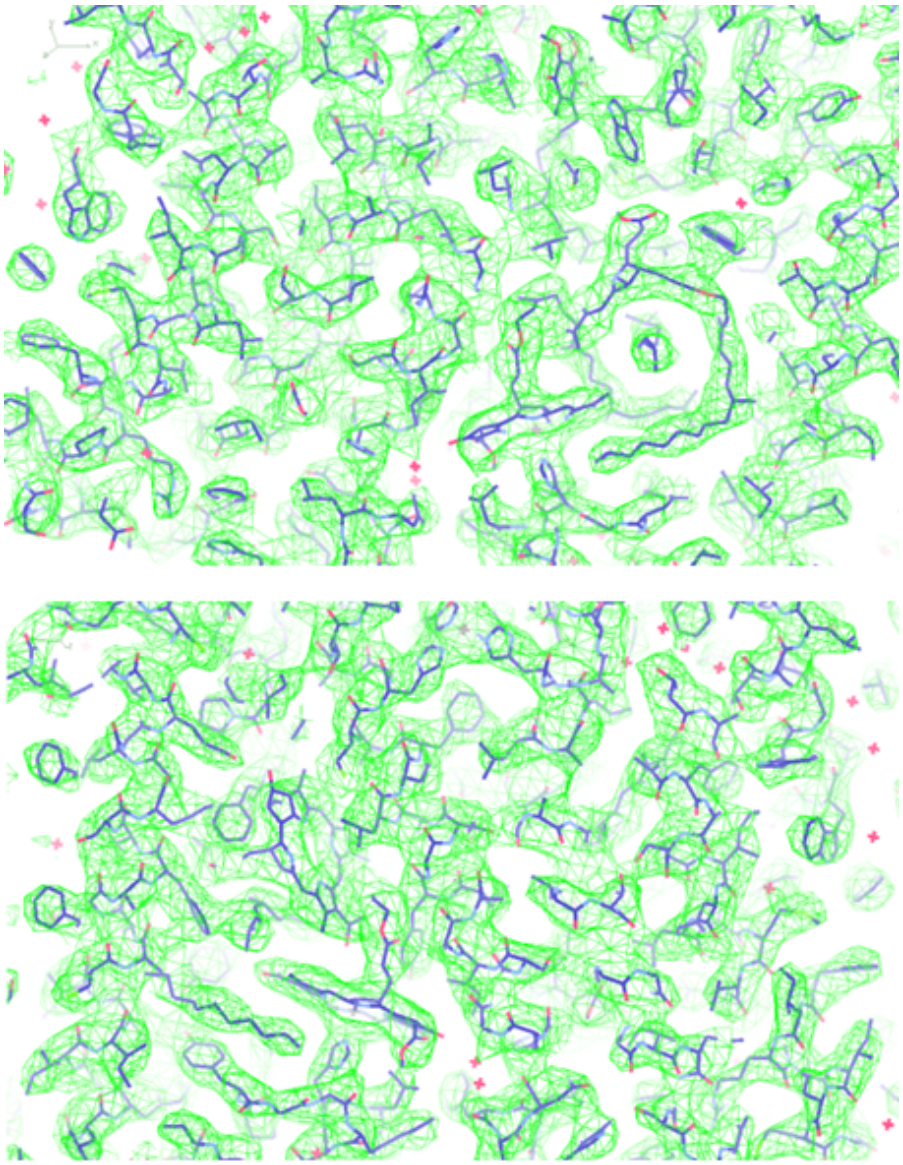

FIGURE 5.8: Typical calculated electron-density maps (green) of 2UXJ at 3.5Å resolution from a successful run in Figure 5.7. The protein structure of 2UXJ has been superimposed (blue and red).

for the protein region has been calculated which includes 40% of the unit-cell volume, equivalent to the volume of the protein plus 8% solvent. Correspondingly, the solvent region computed from the average density map occupies 60% of the unit-cell volume. The calculated results are similar to those of 2UXJ. In particular, the average phase errors are around $30^\circ \sim 40^\circ$. The Fae (1Y5Y) has a fivefold non-crystallographic symmetry,

which was not used in the calculations. Since we have used the density histogram of 1Y5Y to retrieve the phases of 2UXJ, one might expect that the histogram of 2UXJ can be used to retrieve the phases of 1Y5Y. It turns out to be untrue for 2Å data. Instead, the histogram of 1EJB[85] works.

The evolution of phase error and R values is displayed in Figure 5.9. 5% observed reflections are set aside to calculate R_{free}. More than 10, 000 iterations are needed for some runs to converge. The final average phase error is about 38° for 2Å data. When the mean error in phase angle decreases, R_{free} suddenly drops. Because a loose protein mask has been used, R_{work} also has a sudden drop. R values can monitor the mean error in phase angle. Hence they can indicate successful runs.

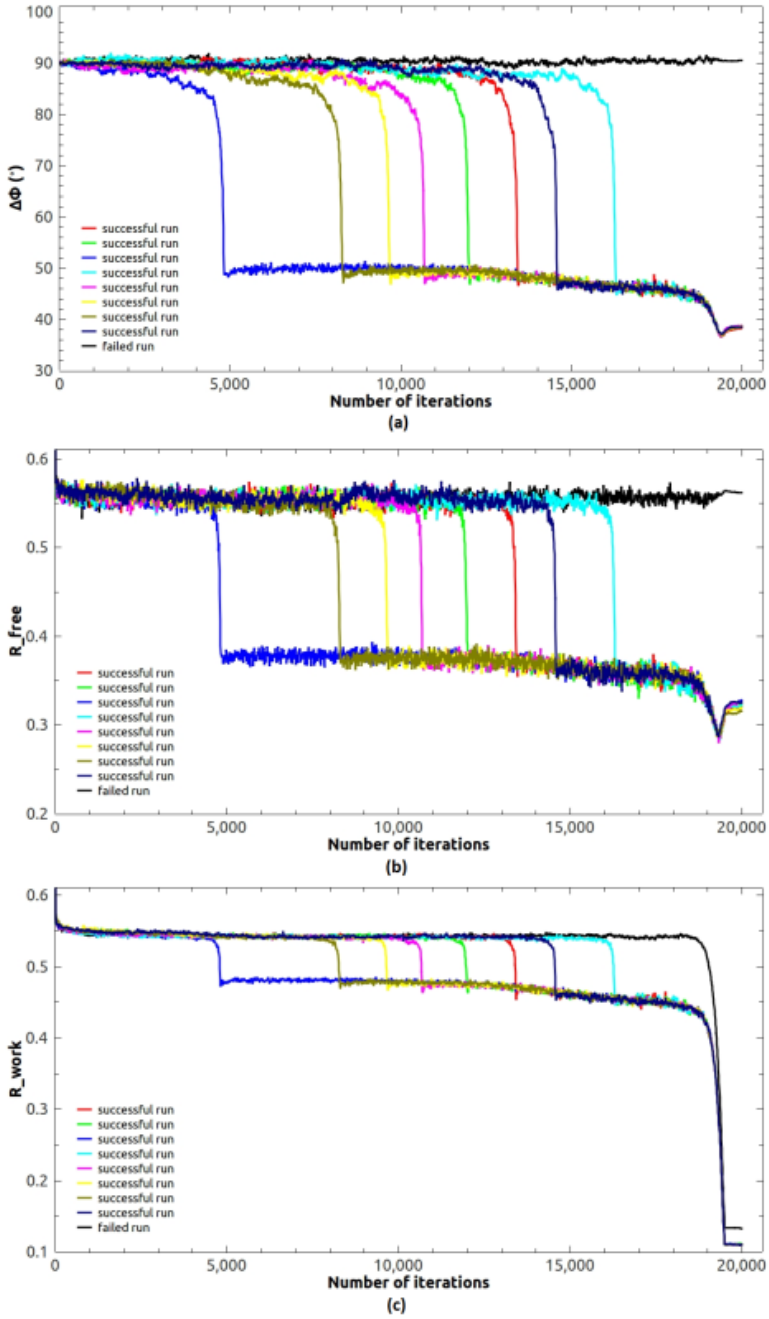

FIGURE 5.9: Evolution of phase error and R values of 2Å data of 1Y5Y.

Typical electron-density maps are shown in Figure 5.10 and the final protein mask is depicted in Figure 5.11. It should be noted that a phase error of 30° can be achieved with the histogram of 1Y5Y itself. Therefore there are probably other structures whose

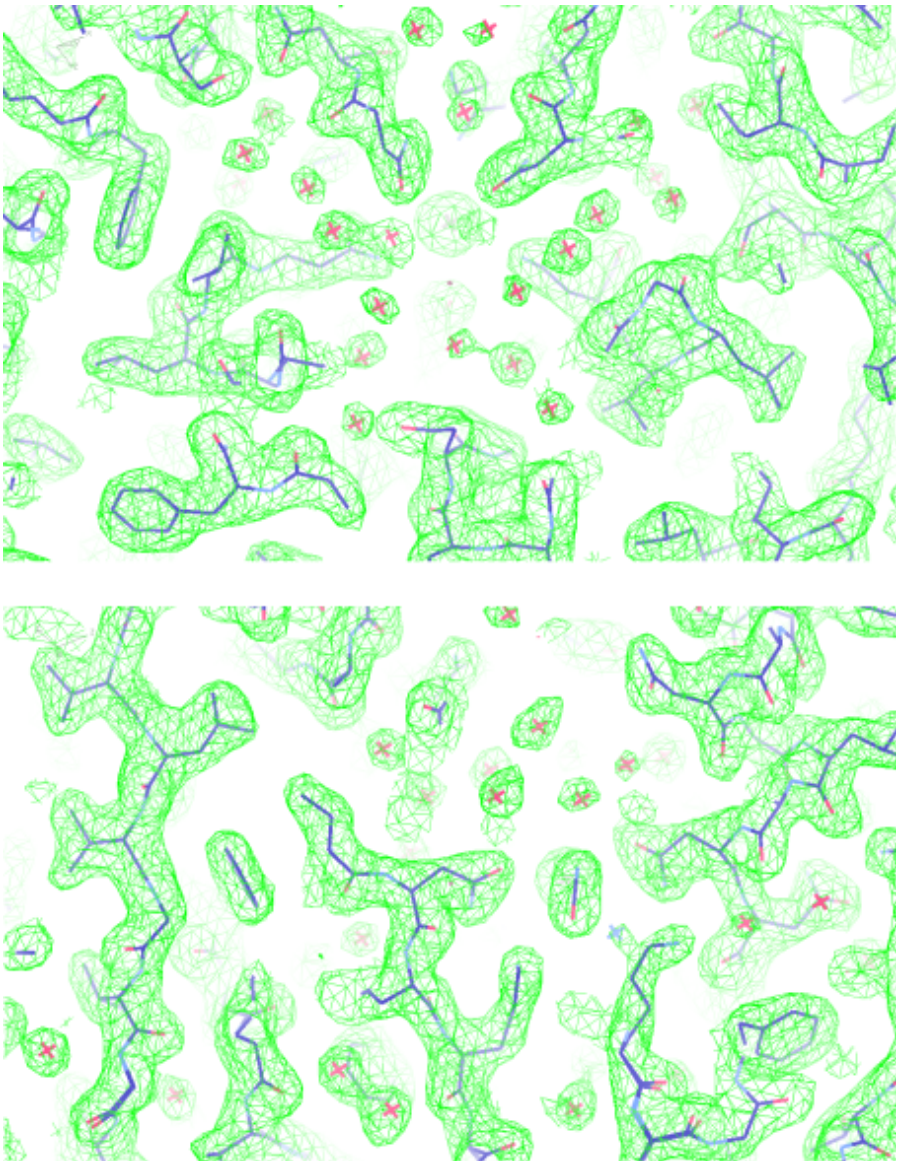

FIGURE 5.10: Typical calculated electron-density maps (green) of 1Y5Y at 2Å resolution from a successful run in Figure 5.9. The protein structure of 1Y5Y has been superimposed (blue and red).

histograms match 1Y5Y better than 1EJB. Another note is that for 1Y5Y, a tight protein mask actually works better than a loose mask in terms of success rate. Finally, at 2.25Å, the histogram of 2UXJ can indeed be used to retrieve the phases of 1Y5Y.

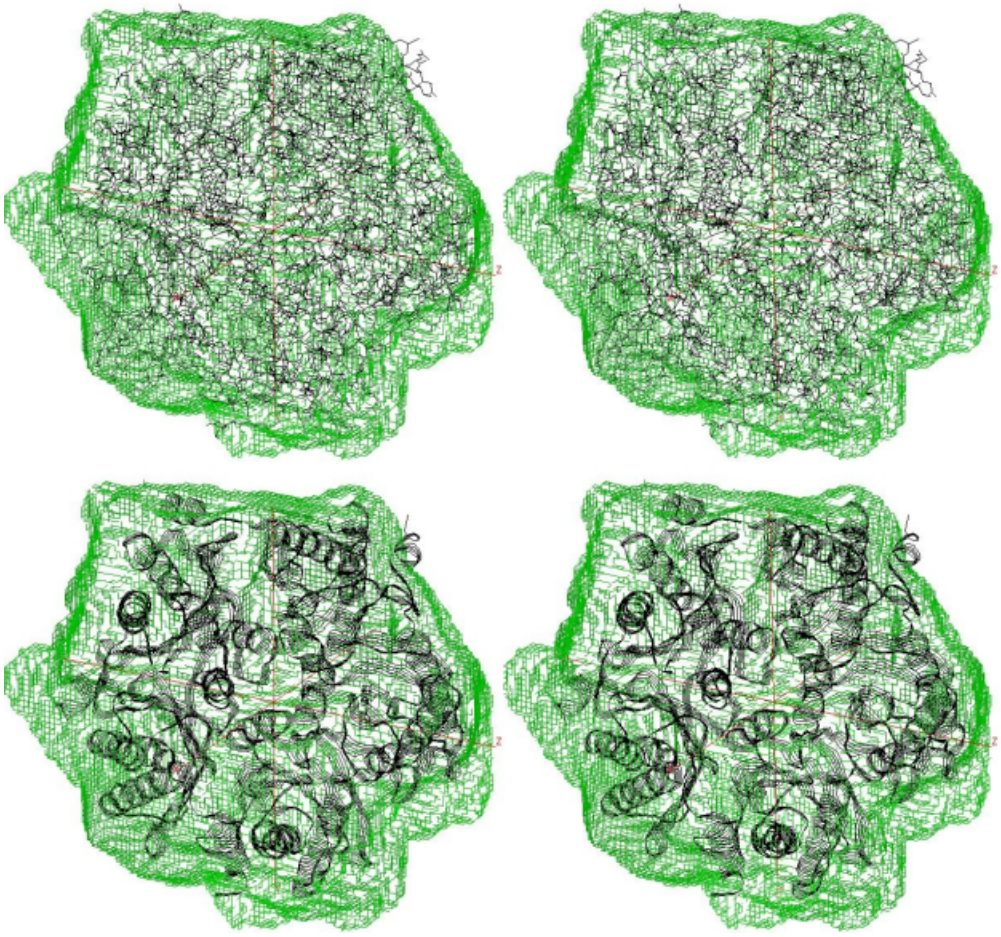

FIGURE 5.11: Stereograms of the calculated protein mask (green) for 1Y5Y from a successful run in Figure 5.9, compared with the atomic model of 1Y5Y in wireframe and strands (black).

The five-fold non-crystallographic symmetry has not been used during the iterative calculations in this chapter. In Chapter 6, we will make use of non-crystallography symmetry and will find NCS averaging can dramatically improve the calculated electron density.

5.4 A human thyroid hormone receptor with PDB ID 3ILZ

The third structure is a human thyroid hormone receptor with PDB ID 3ILZ[72]. The space group is $P2_12_12_1$. The cell dimensions are a = 59.914Å , b = 80.351Å and c = 102.886Å. There are 267 amino acids in the asymmetric unit. The number of non-hydrogen atoms in the asymmetric unit is 2671. The solvent content is 69.43%. The crystal diffracts to 1.85Å , with lowest resolution at 31.47Å. The number of observed unique reflections is 45, 966. The completeness of the data set is 100%, with an overall R value of 0.152. There are only several reflections missing due to the beam stop. However, we find when we use the calculated structure factors to rebuild the reflections below 20Å, our method has better performance. For example the success rate actually increases. The magnitudes of the missing reflections below 1.85Å (including F_{000} and about 47 observed reflections below 20Å) are automatically reconstructed during the iterations. The unit cell is discretized into a $68 \times 92 \times 118$ grid for fast Fourier transform. The distance between two nearby grid points is about 0.9Å. The density function is defined on the grid points. In each iteration cycle, a loose protein mask is calculated which includes 38% of the unit-cell volume which is equivalent to the volume of the protein region plus 7% solvent. Correspondingly, the solvent region computed from the weighted average density map occupies 62% of the unit-cell volume.

The histogram of 1Y5Y at 1.85Å is used as a reference histogram. Histograms from other structures also work sometimes. Histogram matching is only applied to the protein region. In the solvent region, hybrid input-output method is used as a density modification technique.

In order to monitor the calculated results, the mean error in phase angle and R

values are calculated. 5% observed reflections are marked as free data set which is used to calculate R_{free}. In fact, fewer reflections can be used to calculate R_{free}. In that case, the working data set becomes more complete and a high completeness of data always leads to a high success rate. In Figure 5.12, after 20,000 iterations, the mean error in phase angle drops to 50° for successful runs at 1.85Å resolution. When the phase error decreases, a sudden drop of R_{free} occurs. Because a loose protein mask is used, a sudden drop of R_{work} is also observed.

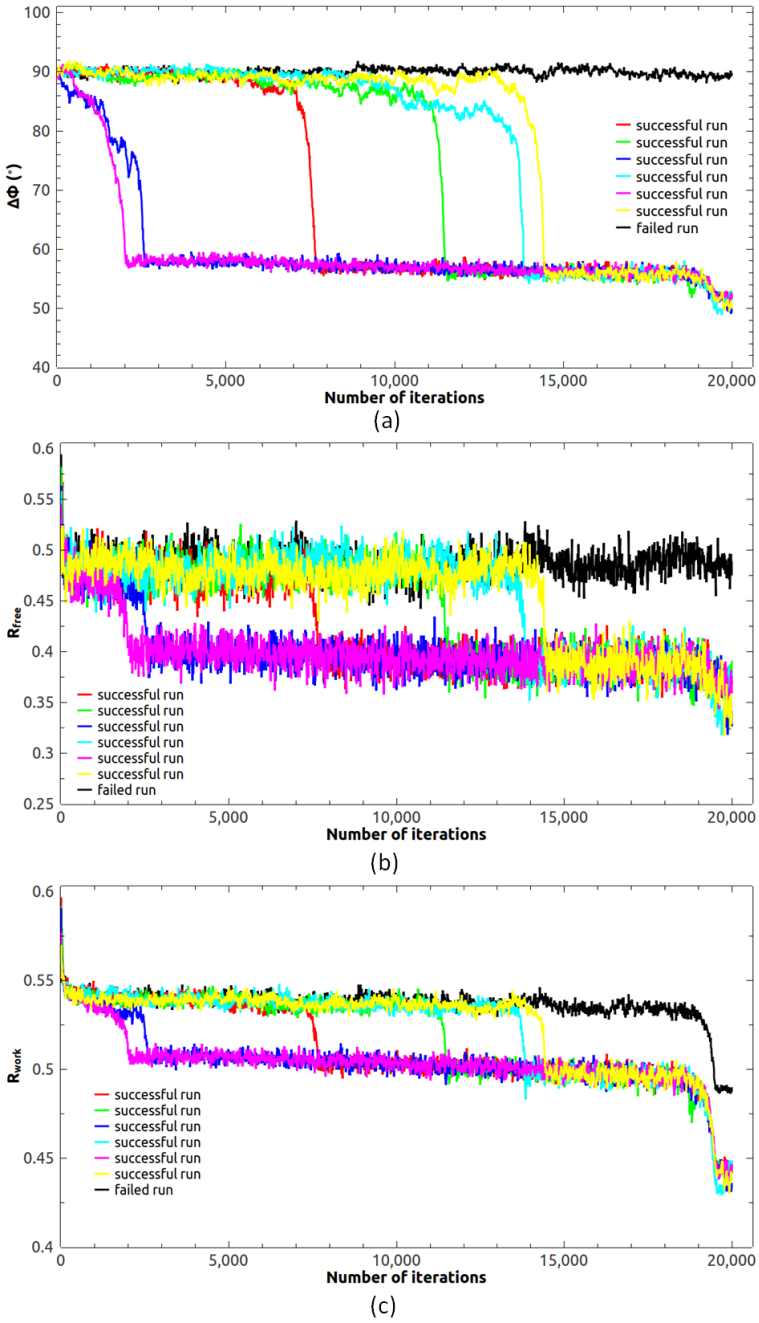

FIGURE 5.12: Evolution of the phase error and R values of 3ILZ.

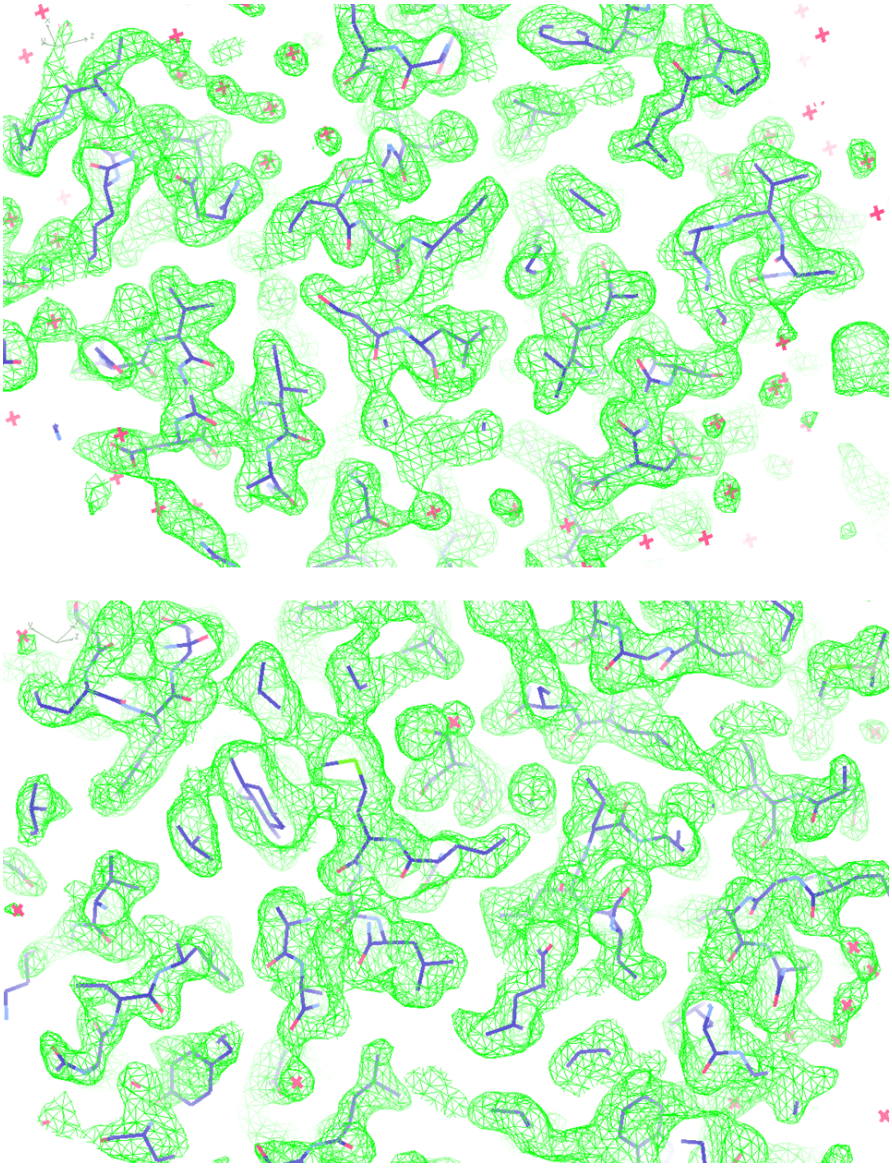

FIGURE 5.13: The calculated 1.85Å electron density maps (green) of 3ILZ from a successful run in Figure 5.12. The protein structure of 3ILZ has been superimposed (blue and red).

For successful runs, the final mean error in phase angle is about 50° for 1.85Å data. If we calculate the mean error in phase angle for 2 Å resolution shell, it is much less than 50°. The calculated electron density is shown in Figure 5.13. It looks very good. The structure of the model can be traced. Most of the fixed water molecules can be located.

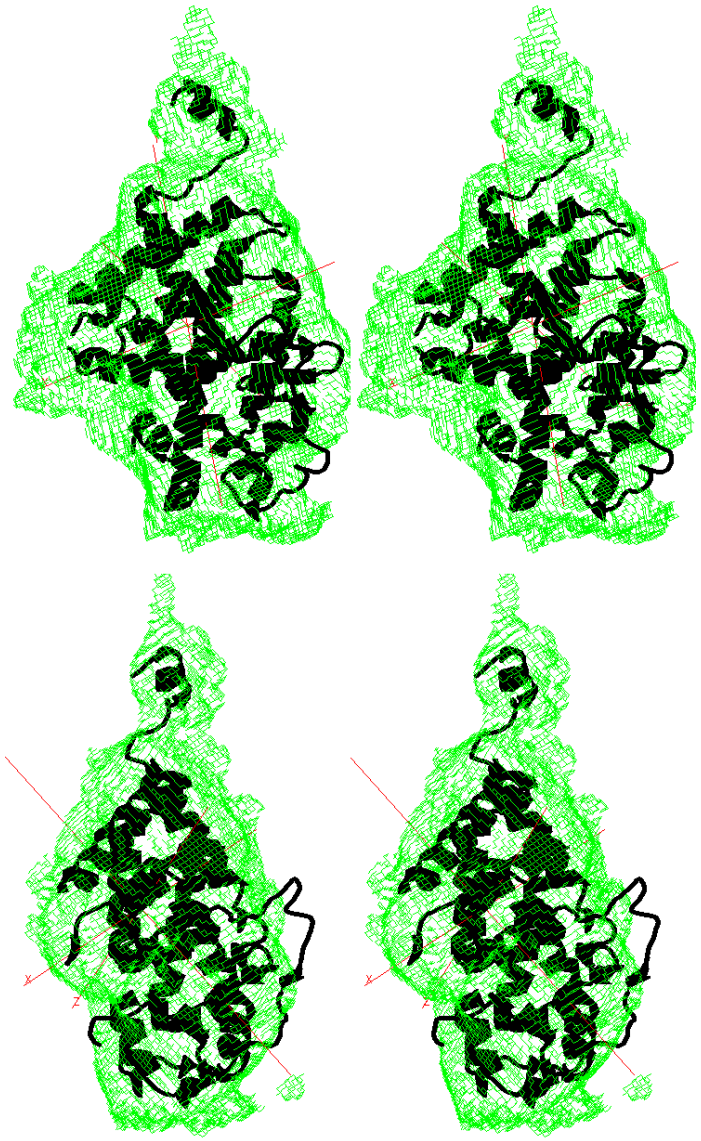

FIGURE 5.14: Stereograms of the calculated protein mask (green) from a successful run in Figure 5.12, compared with the atomic model of 3ILZ in cartoons (black).

The calculated protein mask for a successful run is shown in Figures 5.14 and 5.15. The calculated protein mask can cover most of the protein region, but there are some partial structures which still stay outside of the calculated protein mask. This is normal because the missing reflections can't be exactly rebuilt. The experimental data also contains some measurement errors. When a little big weighting radius is used, the

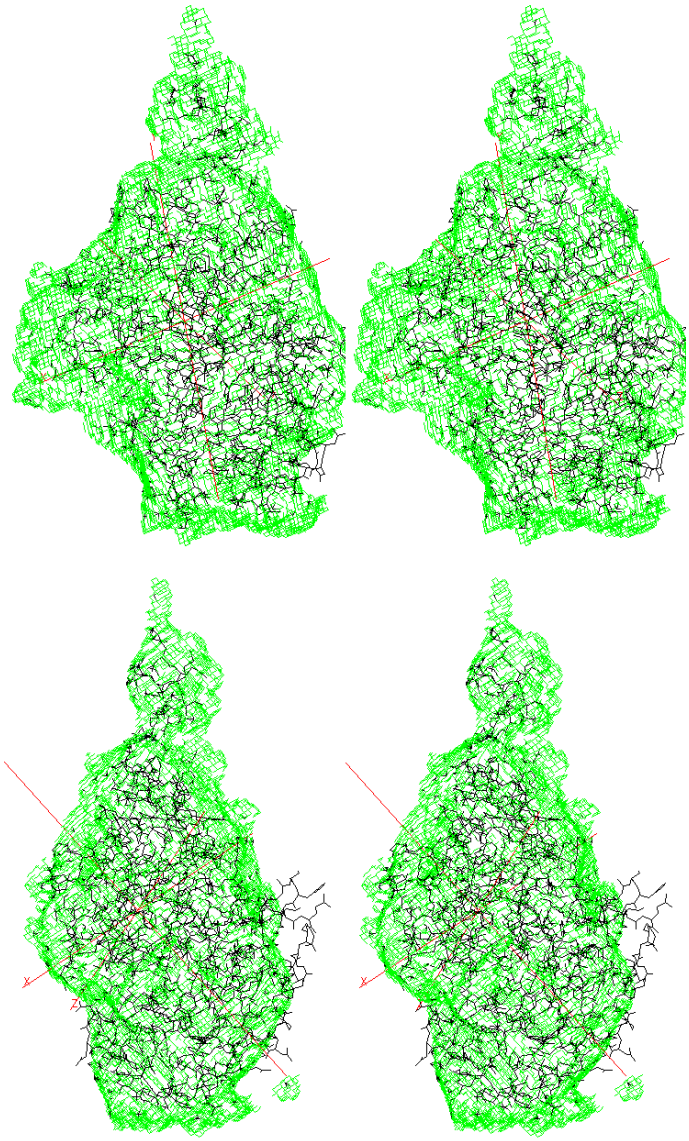

FIGURE 5.15: Stereograms of the calculated protein mask (green) from a successful run in Figure 5.12, compared with the atomic model of 3ILZ in wireframe (black).

resulted protein mask may miss some residues near the protein surface. If most of the protein region can be covered by the calculated protein mask, a successful run can often be reached.

5.5 A pig pancreatic alpha-amylase with PDB ID 1WO2

The fourth structure is a pig pancreatic alpha-amylase with PDB ID 1WO2[84]. The space group is $P2_12_12_1$. The cell dimensions are a = 70.090Å , b = 113.298Å and c = 117.221Å. There are 496 amino acids in the asymmetric unit. The number of non-hydrogen atoms in the asymmetric unit is 4822. The solvent content is 70.00%. The crystal diffracts to 2.01Å , with lowest resolution at 20.00Å. The number of observed unique reflections is 61,363. The completeness of the data set is 98.6%, with an overall R value of 0.159. There are about 61 reflections missing due to the beam stop. The magnitudes of missing reflections below 2.01Å including F_{000} are automatically rebuilt during the iterations. The unit cell is discretized into a $70 \times 114 \times 118$ grid for fast Fourier transform. The distance between two nearby grid points is about 1.0Å. The density function is defined on the grid points. In each iteration cycle, a loose mask for the protein region has been calculated which includes 38% of the unit-cell volume, equivalent to the volume of the protein plus 8% solvent. Correspondingly, the solvent region computed from the average density map occupies 62% of the unit-cell volume.

Several histograms have been tried to serve as the reference histogram but they do not work. At present the histogram of 1WO2 itself is used as a reference histogram. We believe there are many histograms from various structures can be used as the reference histogram.

The mean error in phase angle and R values are used to monitor the calculated result. 5% observed reflections are set aside from the working data set. If fewer reflections are used to calculate R_{free} , the success rate can be higher. The higher the completeness of working data set is, the greater the success rate becomes.

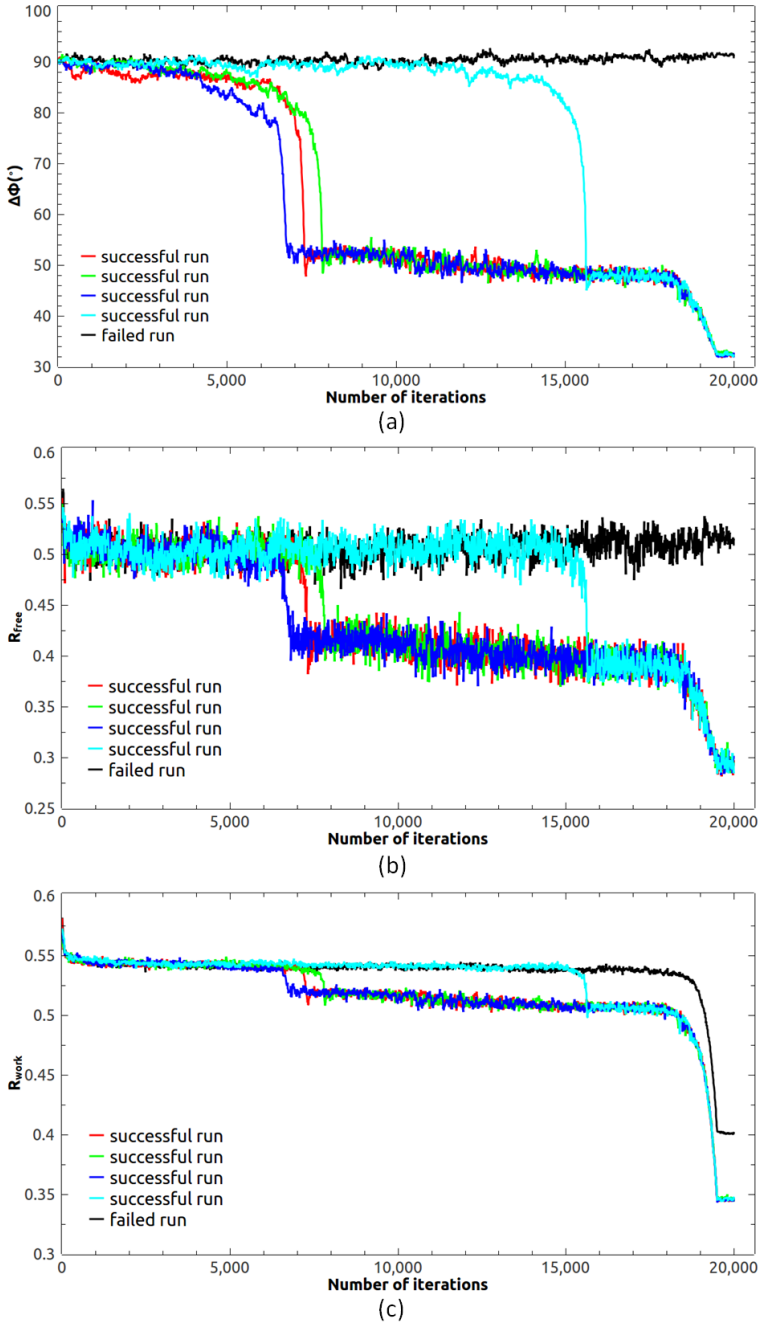

FIGURE 5.16: Evolution of the phase error and R values of 1WO2.

Successful runs have been indicated by sudden drops on the curves of R_{free}. Because a loose protein mask is used, R_{work} also has a sudden drop for a successful run. After 20,000 iterations, the mean error in phase angle drops to 32° for a successful run. The

calculated electron density maps are shown in Figure 5.17. The model structure can be clearly traced and fixed water molecules can be located.

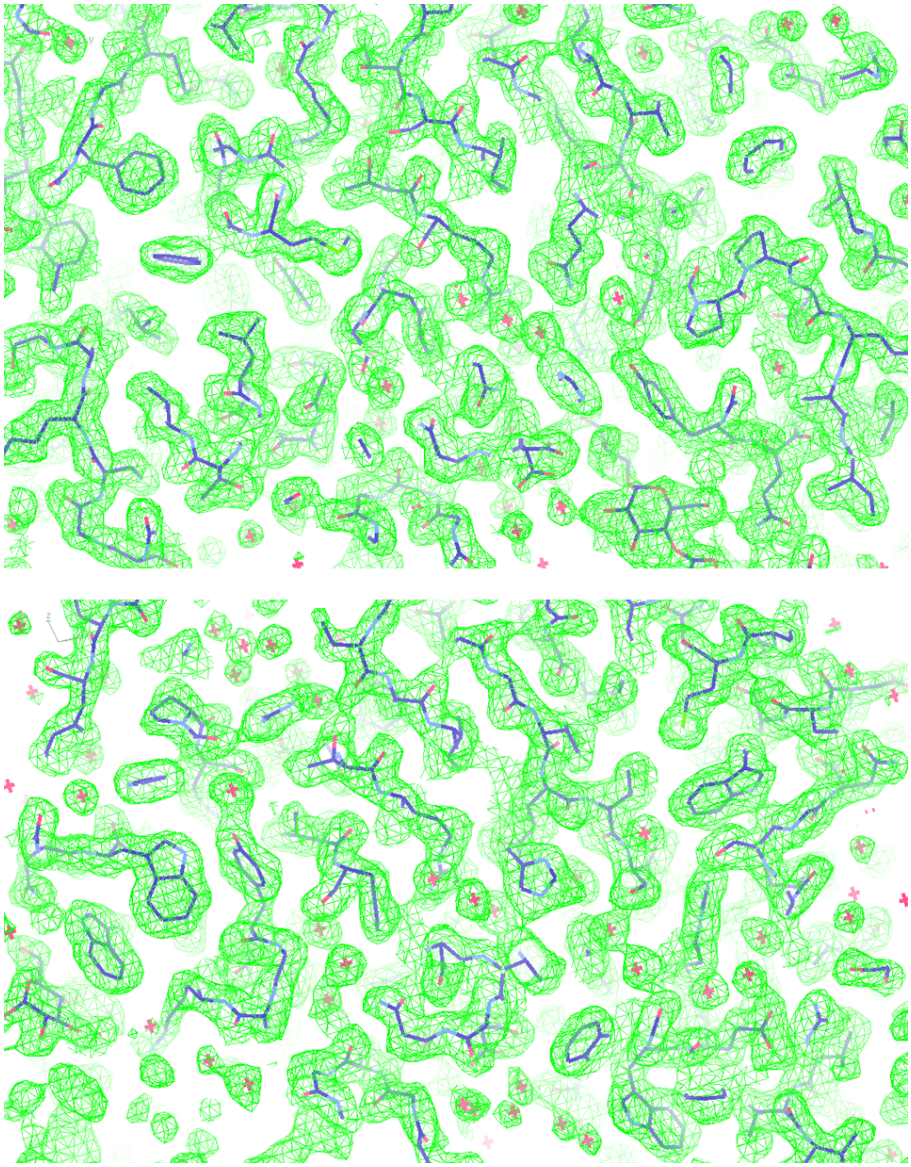

FIGURE 5.17: The calculated 2.01Å electron density maps (green) of 1WO2 from a successful run in Figure 5.16. The protein structure of 1WO2 has been superimposed (blue and red).

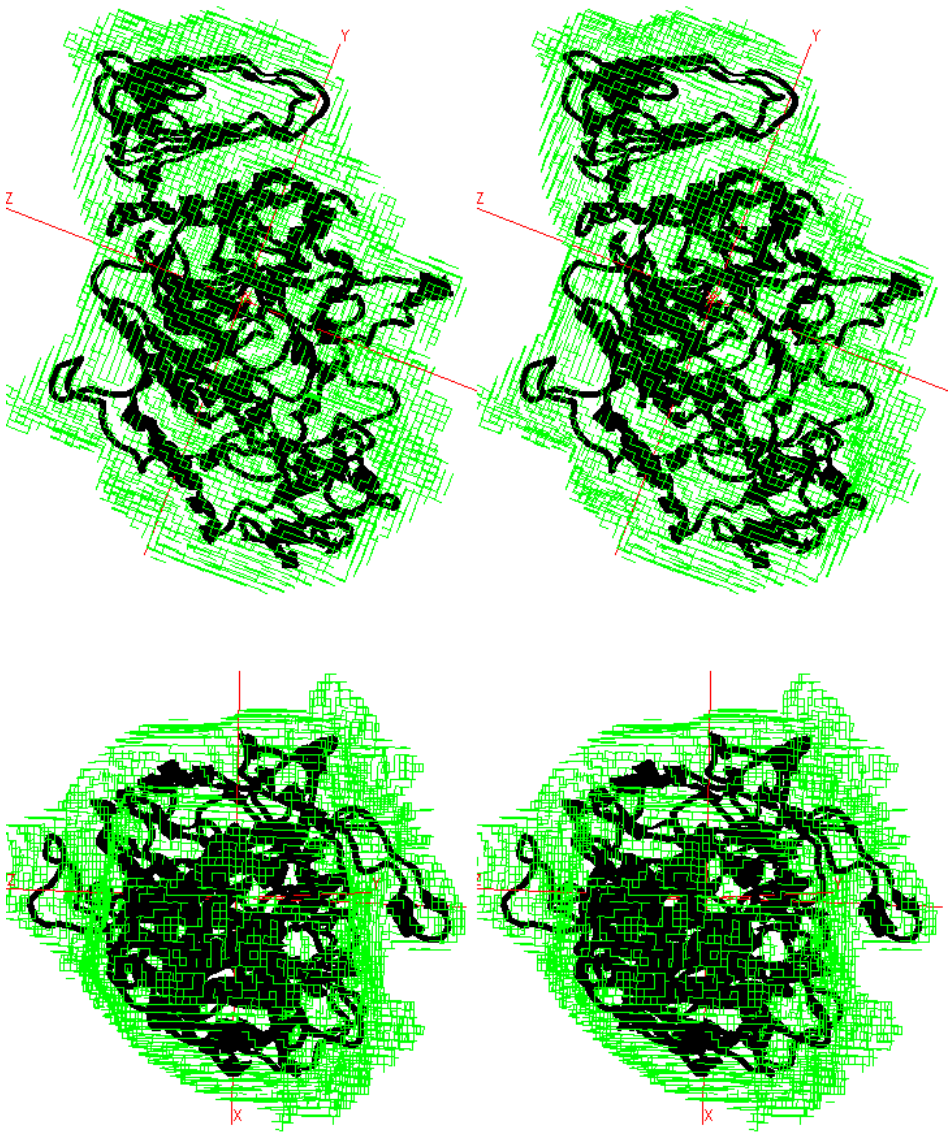

FIGURE 5.18: Stereograms of the calculated protein mask (green) from a successful run in Figure 5.16, compared with the atomic model of 1WO2 in cartoons (black).

The calculated protein mask is shown in Figures 5.18 and 5.19. The calculated protein mask covers the protein molecule very well. Almost all structures on the protein are inside the protein mask. This can explain why the final phase error can be reduced to as small as 32°.

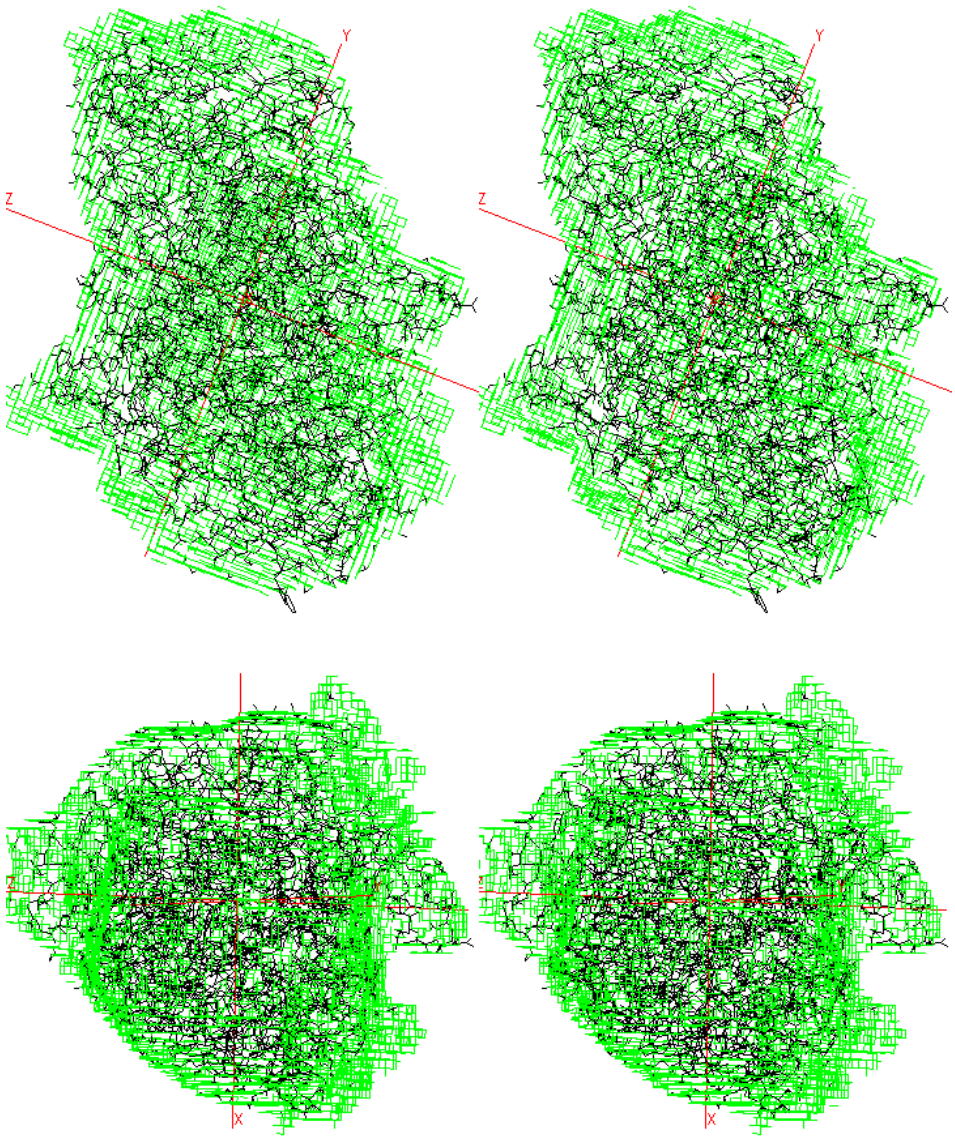

FIGURE 5.19: Stereograms of the calculated protein mask (green) from a successful run in Figure 5.16, compared with the atomic model of 1WO2 in wireframe (black).

5.6 A flavor protein WrbA from *Escherichia coli* with PDB ID 3B6I

The fifth structure is a flavor protein WrbA from *Escherichia coli* with PDB ID 3B6I[83]. The space group is $P4_32_12$. The cell dimensions are a = 94.361Å , b = 94.361Å and c = 175.363Å. The asymmetric unit has D2 non-crystallographic symmetry. There are 396 amino acids in the asymmetric unit. The number of non-hydrogen atoms in the asymmetric unit is 3609. The solvent content is 73.71%. The highest resolution of the diffraction data is 1.66Å and the lowest resolution is 29.41Å. The number of observed unique reflections is 86,954. The completeness of the data set is 97.5%, with an overall R value of 0.170. There are about 31 reflections missing due to the beam stop. The magnitudes of missing reflections below 1.66Å including F_{000} are automatically reconstructed during the iterations. The unit cell is divided into a $114 \times 114 \times 212$ grid for fast Fourier transform, the distance between two nearest grid points is about 0.8Å. The density function is defined on the grid points. In each iteration cycle, a loose mask for the protein region has been calculated which includes 42% of the unit-cell volume, equivalent to the volume of the protein plus 16% solvent. Correspondingly, the solvent region computed from the average density map occupies 58% of the unit-cell volume.

A pretty loose protein mask has been calculated in order to make the calculated protein mask cover the whole protein region. A loose protein mask means smaller solvent region. Our method is based on the oversampling condition. Small solvent region is not good and it always leads to more iterations and low success rate.

The histogram of 3B6I itself has been used to serve as the reference histogram when

histogram matching is applied inside the calculated protein mask. 20,000 iterations are taken on the working data set of 3B6I. Because the completeness of the observed data is not very high, the free data set only contains 1% observed reflections.

The success rate is very low. In order to get a successful run, we let the calculated protein mask evolve slowly during the iterations. For example, in each iteration cycle, only 10% of the previous calculated protein mask is updated. Several schemes can be used to control the update speed of the calculated protein mask. A slowly-updated protein mask gives the calculated density inside the mask more time to evolve to the correct value. However, for many other structures, a fast-updated protein mask means few iteration cycles are needed to reach a successful run.

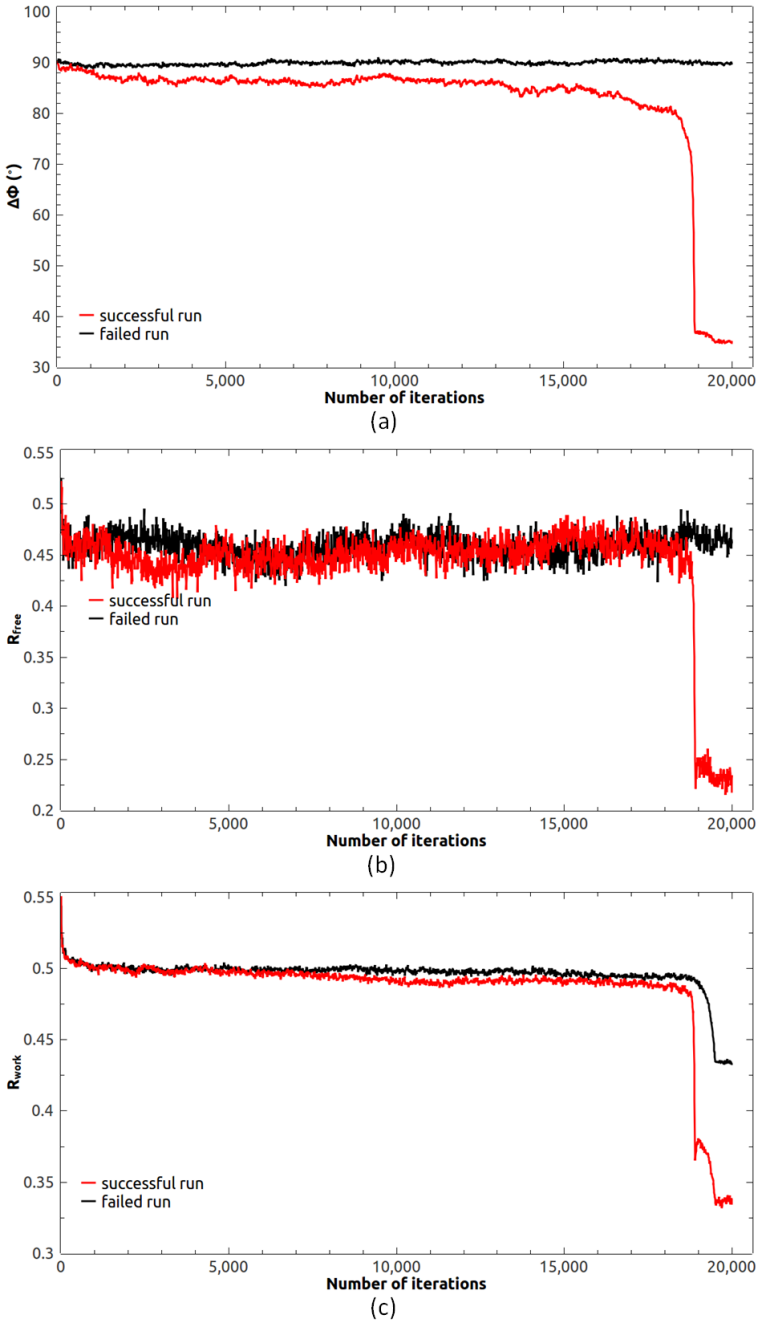

FIGURE 5.20: Evolution of the phase error and R values of 3B6I.

The evolution of the mean error in phase angle and R values is shown in Figure 5.20. A sudden drop of R_{free} indicates a successful run. Because a pretty loose protein mask is used, R_{free} also drops suddenly to indicate a successful run. The final mean error in

phase angle is about 35° for a successful run which leads to very good electron density maps shown in Figure 5.21.

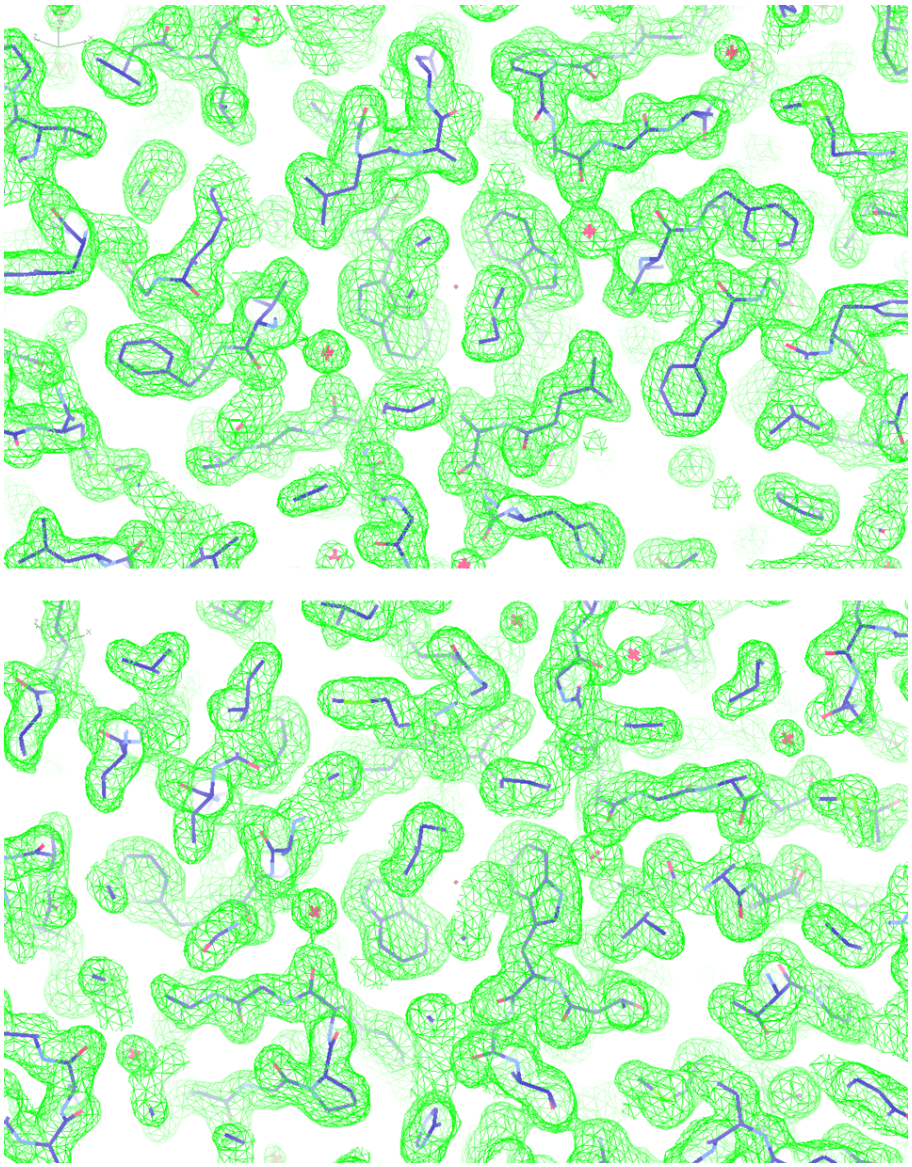

FIGURE 5.21: The calculated 1.66Å electron density maps (green) of 3B6I from a successful run in Figure 5.20. The protein structure of 3B6I has been superimposed (blue and red).

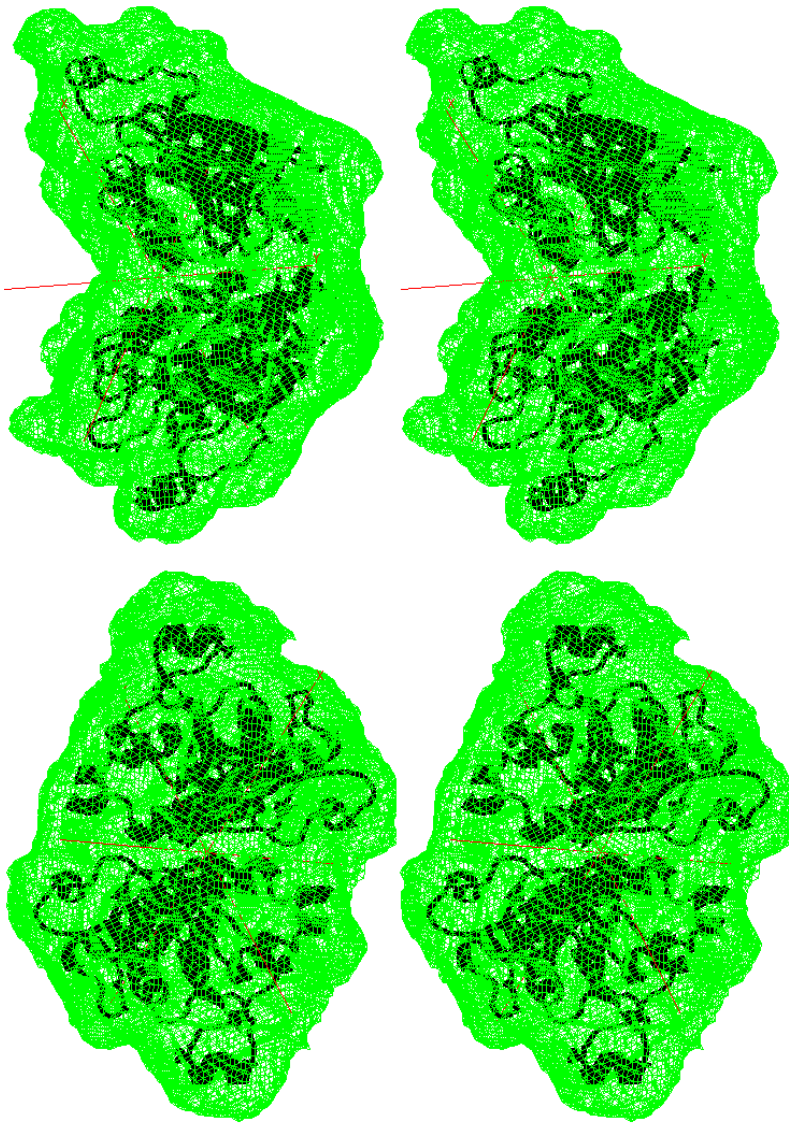

FIGURE 5.22: Stereograms of the calculated protein mask (green) from a successful run in Figure 5.20, compared with the atomic model of 3B6I in cartoons (black).

The calculated protein mask for a successful run is shown in Figures 5.22 and 5.23. The calculated protein mask completely covers the whole protein molecule. Because the protein mask is pretty loose, solvent regions inside the protein mask can be observed.

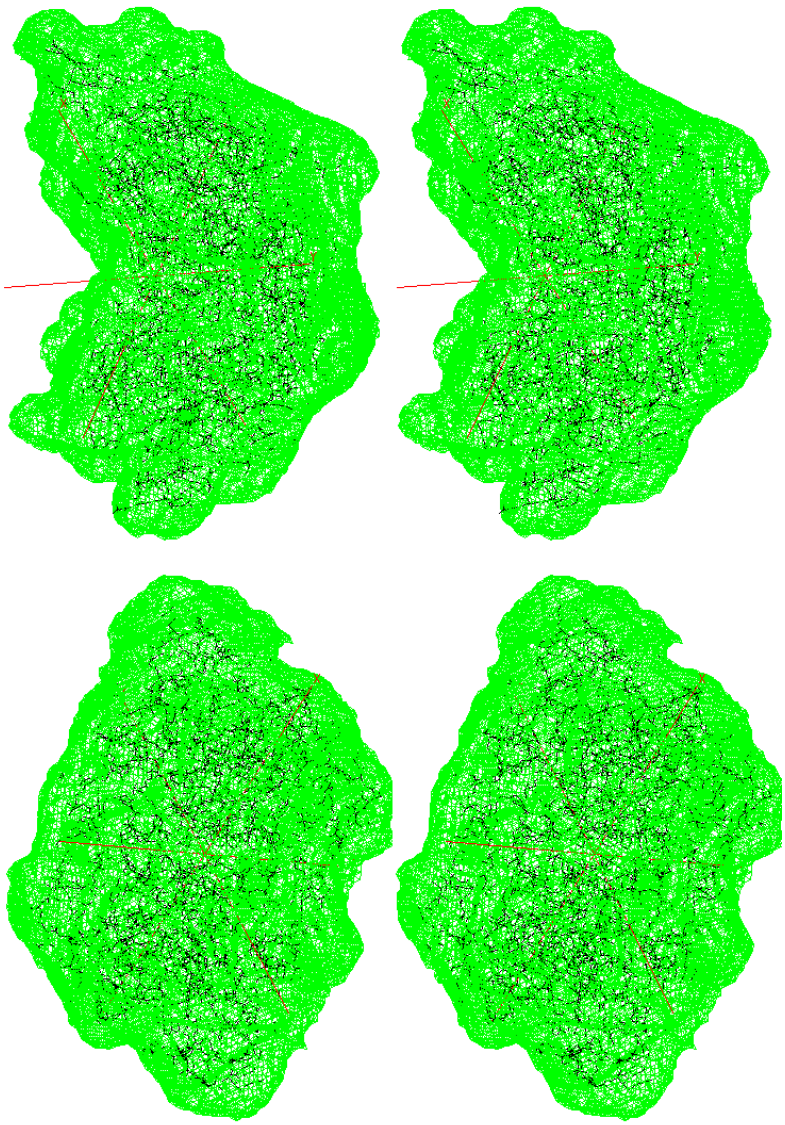

FIGURE 5.23: Stereograms of calculated protein mask (green) from a successful run in Figure 5.20, compared with the atomic model of 3B6I in wireframe (black).

A pretty loose protein boundary is preferred for 3B6I. The solvent content used in the iterations is 58% of the unit-cell volume. The small solvent region challenges oversampling condition. It results in a very low success rate and tens of thousands of iterations. In order to get a successful run within fewer iteration cycles, we use a two-step strategy. The first step is to retrieve a protein mask directly from the diffraction data within several thousand iterations. A good protein mask can be selected if it corresponds to smaller R values. The second step is to keep the pre-calculated protein mask fixed while retrieving the phases. A calculated protein mask is shown in Figures 5.24 and 5.25 after several thousand iterations. There are some small partial structures outside of the protein mask. Although this pre-calculated protein mask is not very good, sometimes, it can make our method find a solution faster which is shown in Figure 5.26. I have to point out that this two-step scheme is not necessary for most crystals especially those with high solvent content. But this two-step strategy indeed speeds up the convergence in most cases.

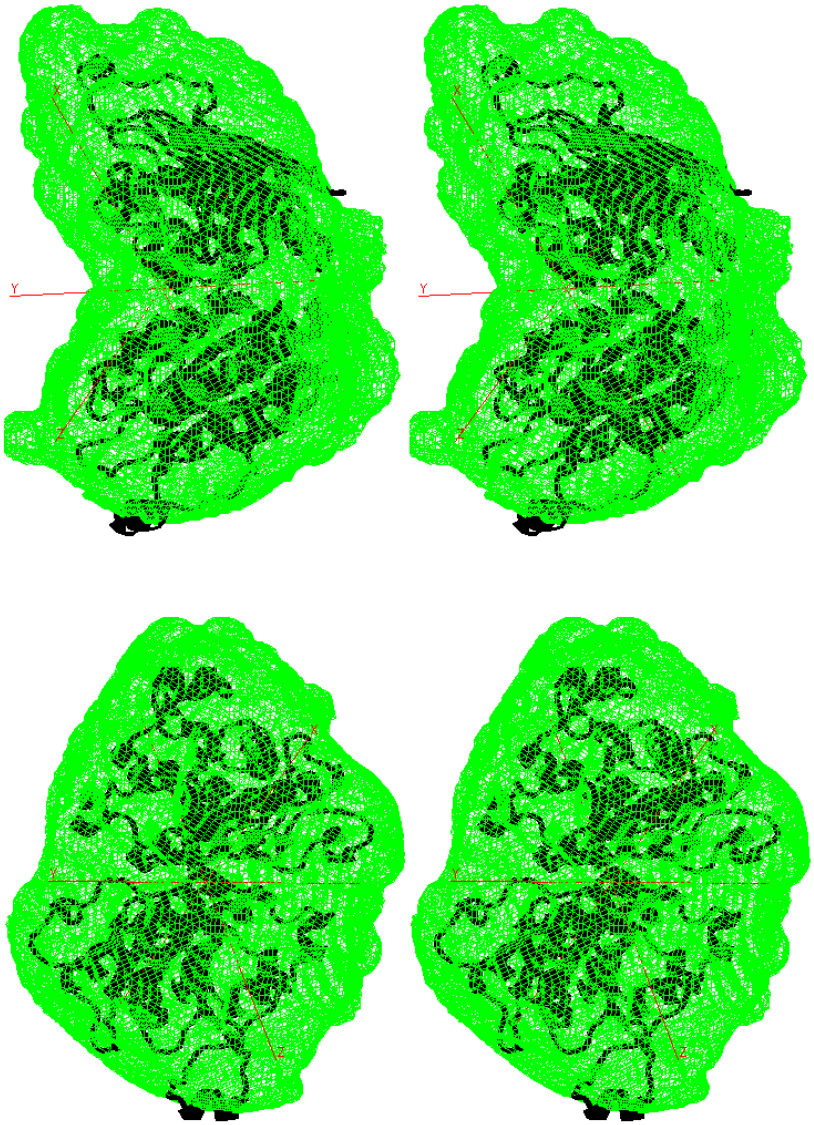

FIGURE 5.24: Stereograms of a pre-calculated protein mask (green) compared with the atomic model of 3B6I in cartoons (black).

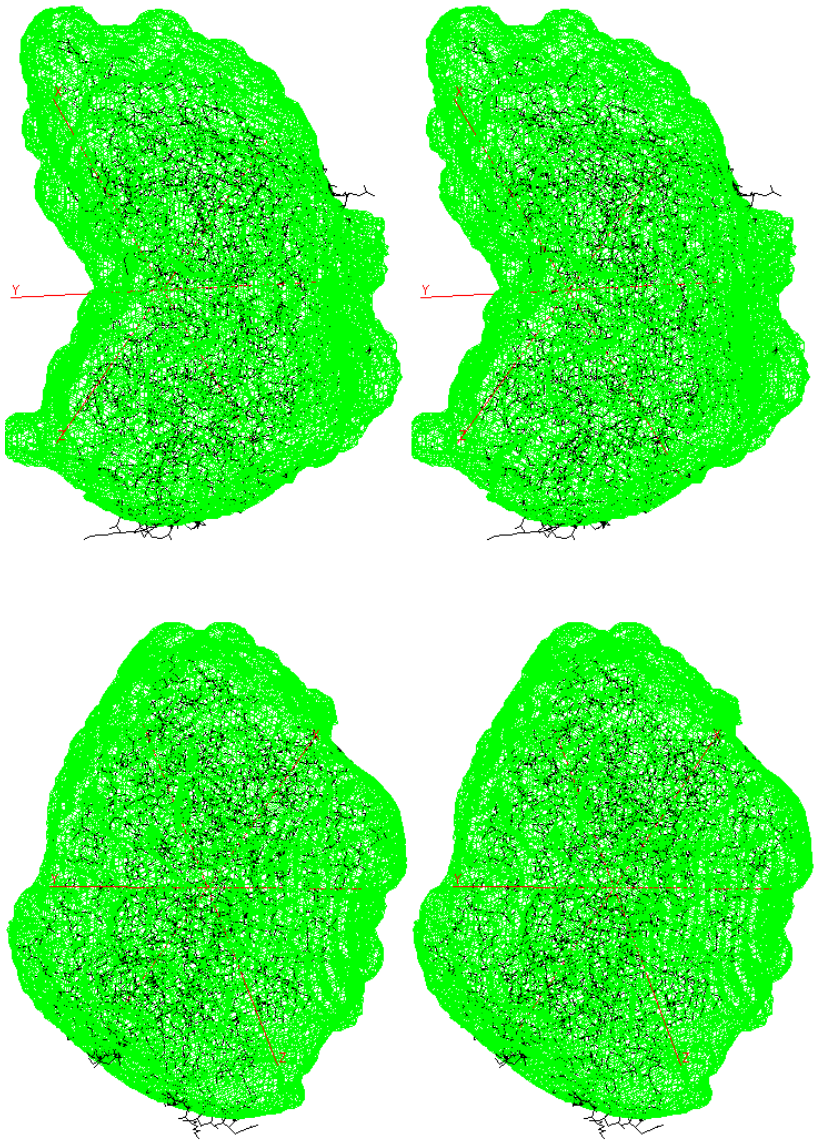

FIGURE 5.25: Stereograms of a pre-calculated protein mask (green) compared with the atomic model of 3B6I in wireframe (black).

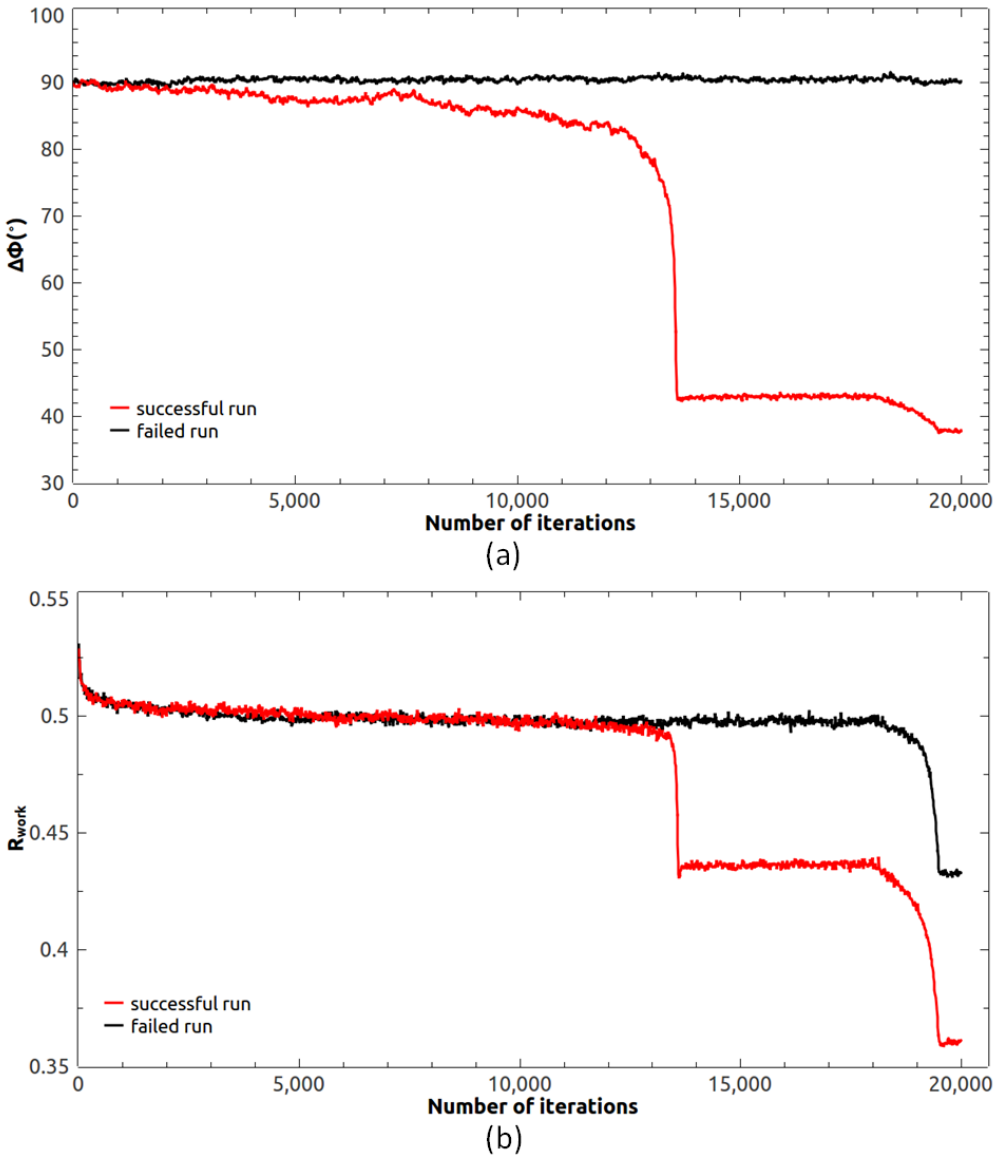

FIGURE 5.26: Evolution of the phase error and R values of 3B6I with a pre-calculated protein mask shown in Figures 5.24 and 5.25.

The D2 non-crystallographic symmetry of the asymmetric unit has not been used during the iterative calculations. In next chapter, non-crystallographic symmetry will be used to calculate the average electron density in the asymmetric unit and the calculated phases will be improved a lot.

5.7 Discussions

For protein crystals with 50% or higher solvent content, Liu et al. have provided strong evidence that phasing through iterative transform is possible provided a protein envelope is available. What we have shown through the trial calculations of 2UXJ and other structures is that the assumption of an envelope is not necessary, and therefore *ab initio* phasing is possible. The generality of our methodology strongly supports the claim that direct phasing is possible for many high-solvent-content protein crystals.

Although our primary concern is direct phasing, the algorithm can be used to supplement and enhance many existing refinement tools. Partial knowledge of some of the phases, for example, can easily be incorporated in the iteration. Prior real-space information such as solvent region [86] or protein fragments can also be employed in the density modification. The large number of iterations helps to eliminate the bias of initial phases.

Although our *ab initio* iterative phasing algorithm resembles the conventional solvent flattening[79] refinement, there are important differences. It is instructive to make a comparison. First, the density modification in the solvent region via the HIO scheme is much more powerful than simply setting the solvent density to a constant. It is well known, for example, that HIO can overcome the stagnation problem and therefore it makes the convergence toward the correct solution possible, whatever the initial starting phases are. Secondly, the number of iterations matters. We have seen in the trial calculations that tens of thousands of iterations are needed in general to retrieve the phases correctly. Thirdly, the missing low-resolution central reflections (due to the beam stop) are not included in the conventional refinement, but they are dominant terms in the

Fourier expansion of the density function and they can greatly affect the construction of the shape[87–89] of the protein. In our phasing scheme, they are reconstructed[63, 90] from the calculated values; thus a very complete Fourier expansion of the density is achieved, and therefore very accurate phases can be retrieved.

Finally, a most important new feature of our algorithm is the evolution of solvent boundary or protein mask. It is true that the boundary could change in the traditional solvent flattening cycles, but not to the extent that happens in our algorithm, where it goes from a completely random boundary to a very accurate one. That is why we can start from random phases and the traditional refinement requires a set of good phases to start with. All of the above factors conspire to make *ab initio* phasing possible. For the same reasons, our iterative scheme can greatly increase the chance of success of a refinement job.

5.8 Conclusions

The traditional way of solving the phase problem starts by collecting the experimentally determined phases, which are rarely accurate enough to yield an interpretable electron density map. Phase improvement using a variety of density modification methods is generally required. Solvent flattening, histogram matching[76, 77], and non-crystallographic averaging[91] are the main techniques. In general, it is believed[92] that density-modification techniques will not turn a bad map into a good one, but they will certainly improve a promising map that shows some interpretable features.

It has gradually been realized[63, 93] in very recent years that a general class of iterative projection algorithms[94, 95], which includes the HIO scheme, can considerably

increase the radius of convergence over the conventional density-modification algorithms. Those algorithms offer the possibility of protein structure determination starting with only information on the molecular envelope[96] and low-order non-crystallographic symmetry.

It turns out that, as we have demonstrated, the convergence region of the iterative projection algorithms can be so large that no prior knowledge of the molecular envelope is needed at all, at least in high-solvent-content crystals. Almost any given initial density or phases will iterate towards the correct density or phases, given a large enough number of iterations. With modest NCS, the same thing can happen for low-solvent-content crystals, shown in next chapter. Thus direct phasing is quite likely for most, if not all, protein crystals.

Chapter 6

Direct phasing of protein crystals with low solvent content

6.1 Introduction

In this chapter, the results described in Section 6.2 have been published[1].

In Chapters 4 and 5, an iterative transform method has been proposed and successfully tested on several protein crystals with high solvent content. Oversampling condition is satisfied when the solvent volume is greater than the protein volume[70].

The solvent content of the protein crystal has a sharp cutoff value at 26%, approximately corresponding to the value for close packed atoms. The most frequent solvent content of protein crystals in Protein Data Bank is about 47%, a little less than 50%. The most frequent solvent contents of nucleic acid crystals and protein-nucleic acid complex

[1]Hongxing He and Wu-Pei Su, Acta Crystallographica Section A: Foundations and Advances, 71(1):92–98, 2015[65]. According to the copyright of that Journal, the article can be re-used in my thesis as long as a full reference to the paper is given.

crystals are about 64% and 60%, respectively[97, 98].

Non-crystallographic symmetry (NCS) arises when there are multiple copies of a molecule within an asymmetric unit. Since this symmetry is local to the asymmetric unit and does not extend to the whole crystal, it is referred to as non-crystallographic symmetry. NCS is common in protein crystals. There are about 1/3 or more protein structures have non-crystallographic symmetry in Protein Data Bank.

NCS enables low-solvent-content crystals to satisfy the oversampling condition. Take the protein crystal with PDB ID 4NF2 as an example. The solvent content is about 45%. This protein has a 3-fold NCS axis. There are three copies of the protein in an asymmetric unit. The independent protein volume is 1/3 of 55%. The solvent volume is much greater than the independent protein volume. Oversampling condition is satisfied.

NCS averaging improves the calculated electron density. NCS copies of the molecule should have the same electron density. Therefore, the average electron density is calculated among different NCS copies.

In addition to NCS, other constraints such as gradient-histogram matching is employed. Like traditional density-histogram matching, the frequency distribution of the gradient of the electron density function is also fairly independent of molecular structures at the same resolution. The gradient histogram of a reference structure is used to scale the gradient histogram of a calculated poor density map. This process improves the poor density map especially when high-resolution data is available.

More density constraints will be employed in the future to reduce the freedom of the protein density. For example, atomicity is another density constraint especially for high-resolution observed data. Sayre's equation can also be applied as a constraint to

improve the calculated phases. The independent protein volume can exceed the solvent volume for low-solvent-content crystals. Oversampling condition becomes satisfied.

This chapter is organized as follows. In Section 2, NCS averaging is described and the calculated results of a protein crystal with NCS are presented. In Section 3, gradient-histogram matching is discussed and the calculated results of a small artificial structure are presented. Conclusions are given in the last section.

6.2 NCS averaging

An important criterion for iterative algorithms such as HIO to work is the requirement that the number of independently measured data points exceeds the number of unknown variables, as first pointed out by Miao et al. (1998)[70]. Thus it is not surprising that with the use of non- crystallographic symmetry (NCS), our method may be extended to phase protein crystals with less than 50% solvent fraction as the NCS reduces that number of unknown variables (the electron density within the protein mask). Liu (2012)[63] has illustrated that by using an artificial structure possessing NCS, assuming that the envelope is available. Millane & Lo (2013)[93] have emphasized the same point. To further demonstrate that possibility, we have studied the structure of a carbamoyltransferase with PDB ID 4NF2 (Center for Structural Genomics of Infectious Diseases, unpublished work). The space group is $P2_12_12_1$. The cell dimensions are a = 85.89 Å, b = 99.89 Å and c = 118.99Å . The asymmetric unit has a C3 non-crystallographic symmetry. The NCS axis is threefold. There are 1020 amino acids in the asymmetric unit. The number of non-hydrogen atoms in the asymmetric unit is 8820. The solvent fraction is 44.79%. The resolution range of the diffraction data extends from 29.23 to 1.74Å. The number of

observed unique reflections is 104,782. The completeness of the observed data is 99.4%. The R value for all observed reflections is 0.147.

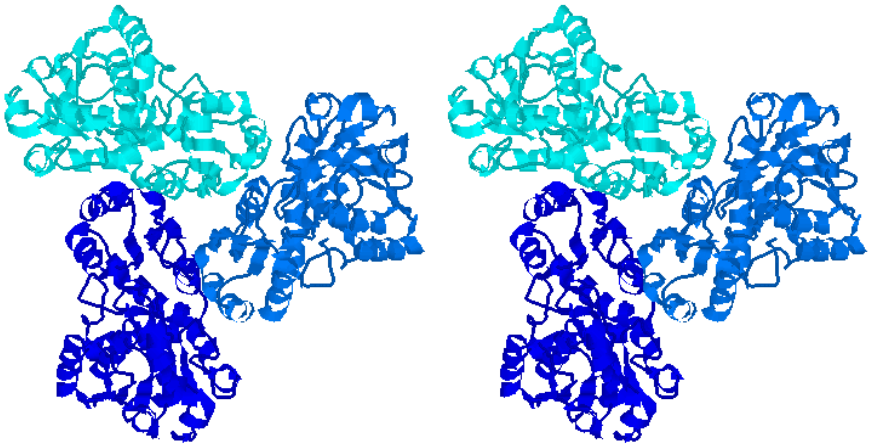

FIGURE 6.1: Stereograms of 4NF2 with C3 non-crystallographic symmetry. There are three copies of the molecule in an asymmetric unit.

As a first step in showing the possibility of an iterative phasing scheme, we assume a given low-resolution envelope (calculated from phases at 30 Å resolution) and the orientation and position of the threefold NCS axis. Synthetic diffraction data are used instead of real data, but bulk solvent correction is taken into account. With those we have carried out HIO iterations with the 1.74 Å data. A reference density histogram was computed from 4NF2 itself at 1.74 Å resolution. Starting essentially from random phases, the evolutions of the phase error at three resolution levels are shown in Figure 6.2. The protein mask is kept fixed throughout the iterations, and the threefold symmetry inside the protein envelope is enforced by conventional NCS averaging in updating the density function. It is clear from Figure 6.2 that correct phases are retrieved after many iterations despite the low solvent content.

The initial phases are almost random which has a mean error about 90° shown in Figure 6.2. At 3Å resolution, the initial mean error in phase angle is a little less than

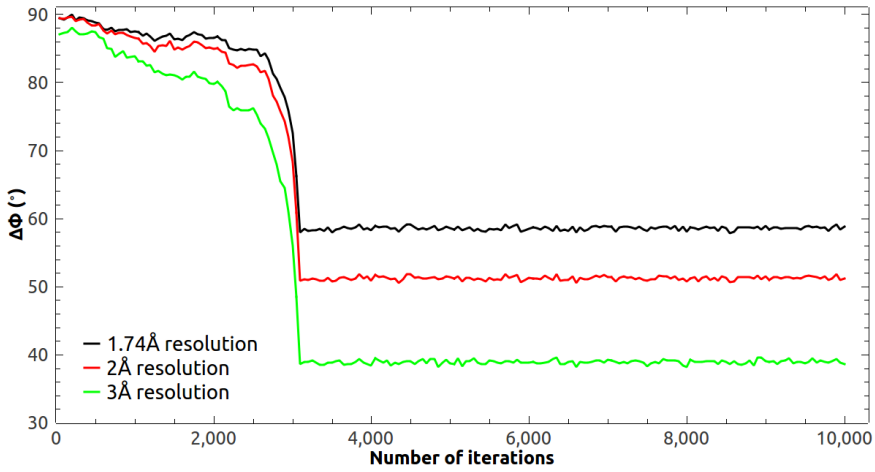

FIGURE 6.2: Evolution of phase error of several resolution ranges of 4NF2 with a fixed protein mask. Synthetic data with a bulk solvent correction has been used.

90° because an initial low-resolution protein mask has been used and fixed during the iterations. After about 3000 iteration cycles, the true phases have been successfully retrieved from the synthetic data. Phases corresponding to lower-resolution reflections become correct ahead of those corresponding to high-resolution reflections. For example, the green line begins to drop ahead of the red and the black lines.

The calculated phases of a successful run are presented in Figure 6.2. The final mean error in phase angle at 1.74 Å resolution is about 58° and the corresponding mean error at 3Å resolution is about 39° . The calculated density maps are shown in Figure 6.3. The density maps look very good due to high resolution data and small mean error in phase angle. The protein structure can be traced on the density map and fixed water molecules can also be located.

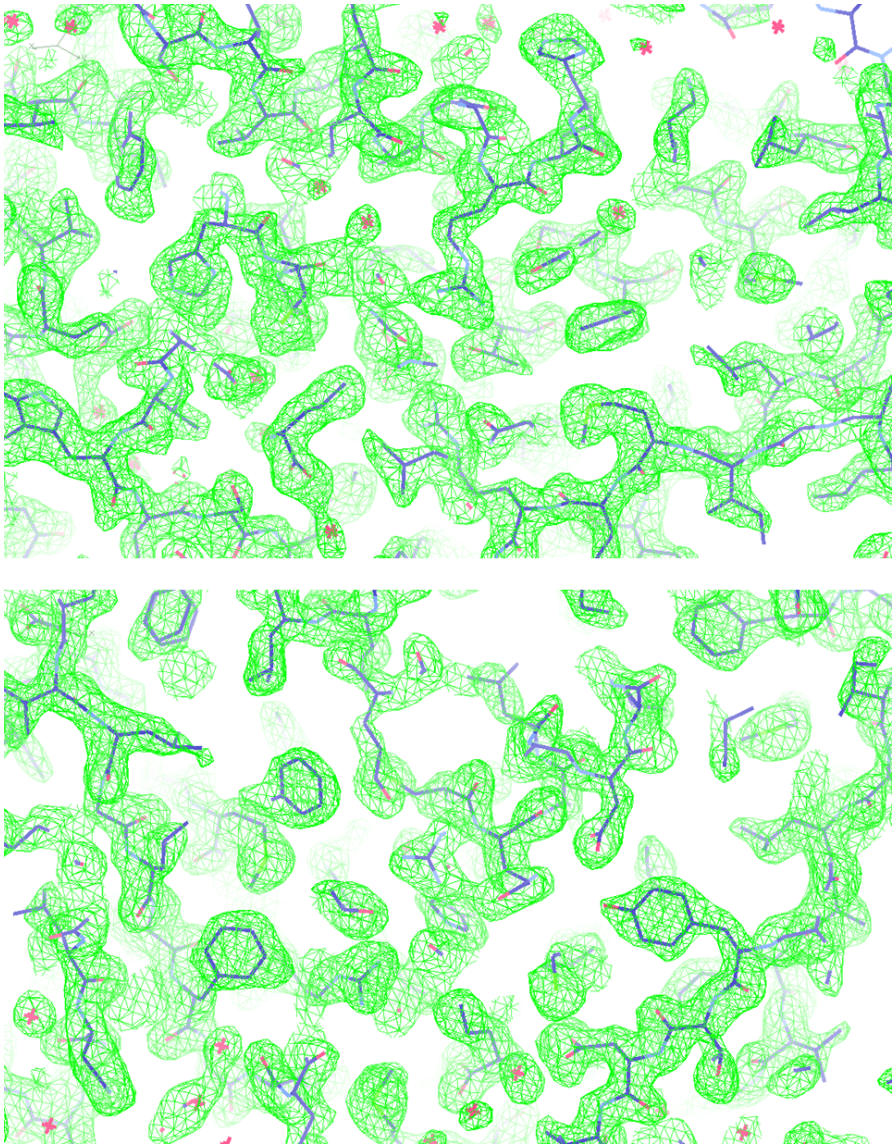

FIGURE 6.3: Calculated density maps (green) of 4NF2.The protein structure of 4NF2 has been superimposed (blue and red).

An approximate NCS mask has been used during the iterative calculations. An exact NCS mask is always preferred. In practice, it is not easy to obtain an exact NCS mask directly from the observed data starting from random phases. The NCS mask used in our calculation is computed from low-resolution data only. The mask is shown in Figures 6.4 and 6.5. There are some small partial structures near the protein surface

sticking out of the NCS mask.

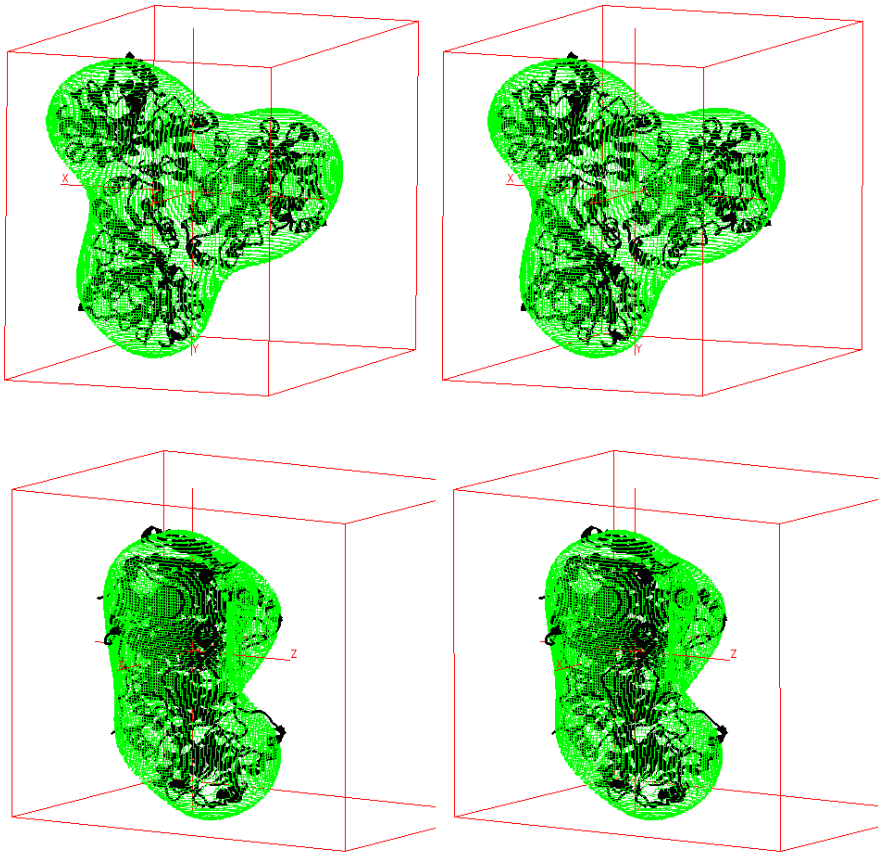

FIGURE 6.4: Stereograms of NCS mask (green) compared with the atomic model of 4NF2 in cartoons (black). The NCS mask is supposed to be known and it is fixed during iterations.

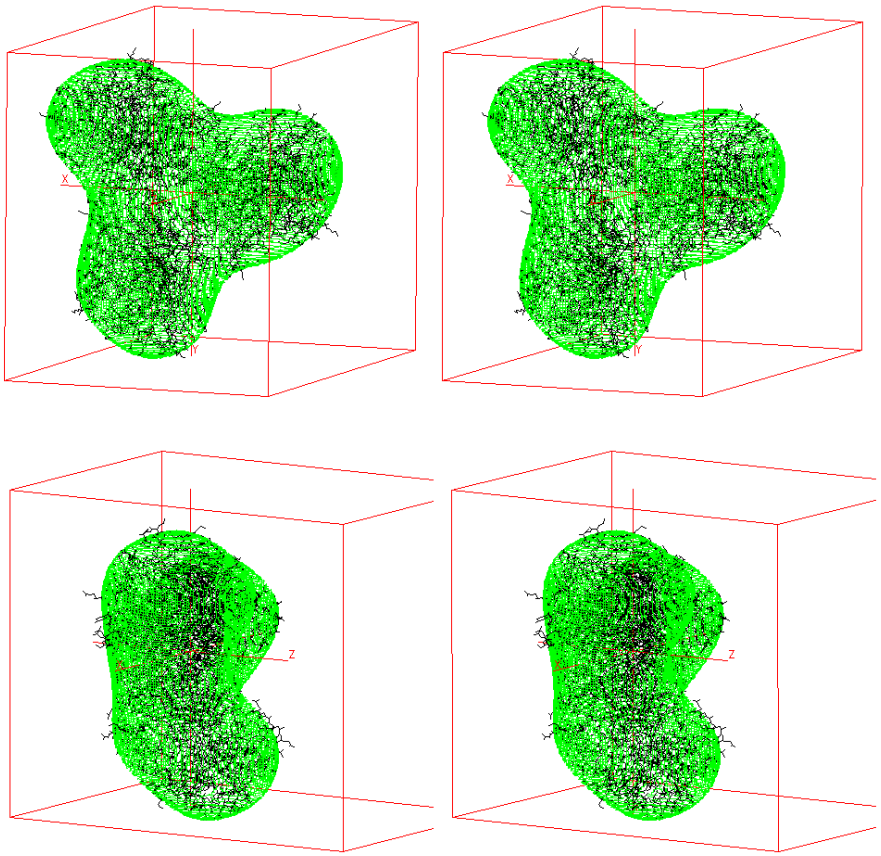

FIGURE 6.5: Stereograms of the NCS mask (green) compared with the atomic model of 4NF2 in wireframe (black). The NCS mask is supposed to be known and it is fixed during iterations.

The protein mask is also given and fixed from the beginning of iterations. The protein mask is calculated at 30Å resolution. It almost exactly overlaps with the NCS mask. In other words, the NCS mask shown in Figures 6.4 and 6.5 is used as the protein mask.

In our test, we have supposed the orientation of the NCS axis is known. The orientation of the NCS axis can generally be found by the self-rotation Patterson map and the native Patterson map. It is not clear how to calculate those entities progressively, unlike the solvent boundary.

6.3 Gradient-histogram matching

The electron density in protein region has similar distributions for different proteins at the same resolution. Similarly, the gradient of the density in the protein region also has similar distributions for different proteins at the same resolution. Gradient matching can smooth the density in the protein region. Both density histogram and gradient histogram are matched at the same time by a joint distribution, referred as two-dimensional or 2D histogram[99, 100].

The experimental data always contain some missing reflections at very low resolution with small diffraction angles. Aside from the missing reflections, measurement error and thermal noise also exist in the observed data. Those factors can make the phase problem complicated. In order to simplify the phase problem and to easily test our method, the synthetic data is a better choice at the beginning.

In this section, synthetic data has been computed directly from the atomic model without any bulk solvent correction. The synthetic data is complete in all resolution shells. If our method works on the synthetic data, with some modifications and improvements, it should work on the real experimental data.

Gradient matching needs a standard gradient histogram. Gradient of the density can be calculated numerically from the electron density map. For example, the density difference in one angstrom gives the gradient at that location. However, this method can only give approximate results. It can't give the exact gradient on a grid point. An analytical method should be exploited to calculate the gradient. The Fourier transform of the electron density yields the structure factors. The gradient of the electron density can also be calculated in the Fourier space.

Gradient is a vector which has three components. Each component can be separately evaluated in the Fourier space. Through an inverse Fourier transform, the three components can be obtained in real space. The modulus of the gradient in real space can also be obtained.

$$\rho(x,y,z) = \frac{1}{V}\sum_{h,k,l=-\infty}^{\infty} F(h,k,l)e^{-2\pi i(hx+ky+lz)} \tag{6.1}$$

$$g_x = \frac{\partial\rho(x,y,z)}{\partial x} = -\frac{2\pi i}{V}\sum_{h,k,l=-\infty}^{\infty} hF(h,k,l)e^{-2\pi i(hx+ky+lz)} \tag{6.2}$$

$$g_y = \frac{\partial\rho(x,y,z)}{\partial y} = -\frac{2\pi i}{V}\sum_{h,k,l=-\infty}^{\infty} kF(h,k,l)e^{-2\pi i(hx+ky+lz)} \tag{6.3}$$

$$g_z = \frac{\partial\rho(x,y,z)}{\partial z} = -\frac{2\pi i}{V}\sum_{h,k,l=-\infty}^{\infty} lF(h,k,l)e^{-2\pi i(hx+ky+lz)} \tag{6.4}$$

which can be inverted as follows.

$$hF(h,k,l) = -\frac{V}{2\pi iN}\sum_{x,y,z} g_x e^{2\pi i(hx+ky+lz)} \tag{6.5}$$

$$kF(h,k,l) = -\frac{V}{2\pi iN}\sum_{x,y,z} g_y e^{2\pi i(hx+ky+lz)} \tag{6.6}$$

$$lF(h,k,l) = -\frac{V}{2\pi iN}\sum_{x,y,z} g_z e^{2\pi i(hx+ky+lz)} \tag{6.7}$$

where V is the volume of the unit cell and N is the total number of grid points in the

unit cell. The summation is over all grid points in the unit cell. g_x, g_y and g_z, are in fractional coordinates. In practice, gradient-histogram matching is often done in the general orthogonal coordinates. g_x, g_y and g_z should be replaced by g_u, g_v and g_w, respectively, in the general orthogonal coordinates. Because the coordinate is in the denominator, gradient transform uses the transpose of the inverse matrix.

$$G_{orth-to-frac} = [M_{frac-to-orth}]^T = \begin{bmatrix} a & 0 & 0 \\ b\cos\gamma & b\sin\gamma & 0 \\ c\cos\beta & \dfrac{c(\cos\alpha - \cos\beta\cos\gamma)}{\sin\gamma} & \dfrac{c\mu}{\sin\gamma} \end{bmatrix} \tag{6.8}$$

$$G_{frac-to-orth} = [M_{orth-to-frac}]^T = \begin{bmatrix} 1/a & 0 & 0 \\ -\dfrac{\cos\gamma}{a\sin\gamma} & \dfrac{1}{b\sin\gamma} & 0 \\ \dfrac{\cos\alpha\cos\gamma - \cos\beta}{a\mu\sin\gamma} & -\dfrac{\cos\alpha - \cos\beta\cos\gamma}{b\mu\sin\gamma} & \dfrac{\sin\gamma}{c\mu} \end{bmatrix} \tag{6.9}$$

Although the electron density has crystallographic symmetry, the gradient of the density does not hold the crystallographic symmetry. In other words, the gradient does not have any symmetry operations or equivalent positions and it should be calculated throughout the whole unit cell.

Gradient histogram contains three component histograms and one modulus histogram. Firstly, gradient-histogram matching should be applied in the x, y and z directions, respectively. Secondly, the modulus of the gradient should also be matched. When the gradient-modulus matching is applied, three gradient components should be

modified with the same scale. After gradient-modulus matching, the three gradient components don't need to be matched again.

A small artificial structure has been made to test 2D histogram matching inside the calculated protein mask. This artificial structure looks like a sphere and consists of three short alpha helices, as shown in Figure 6.6.

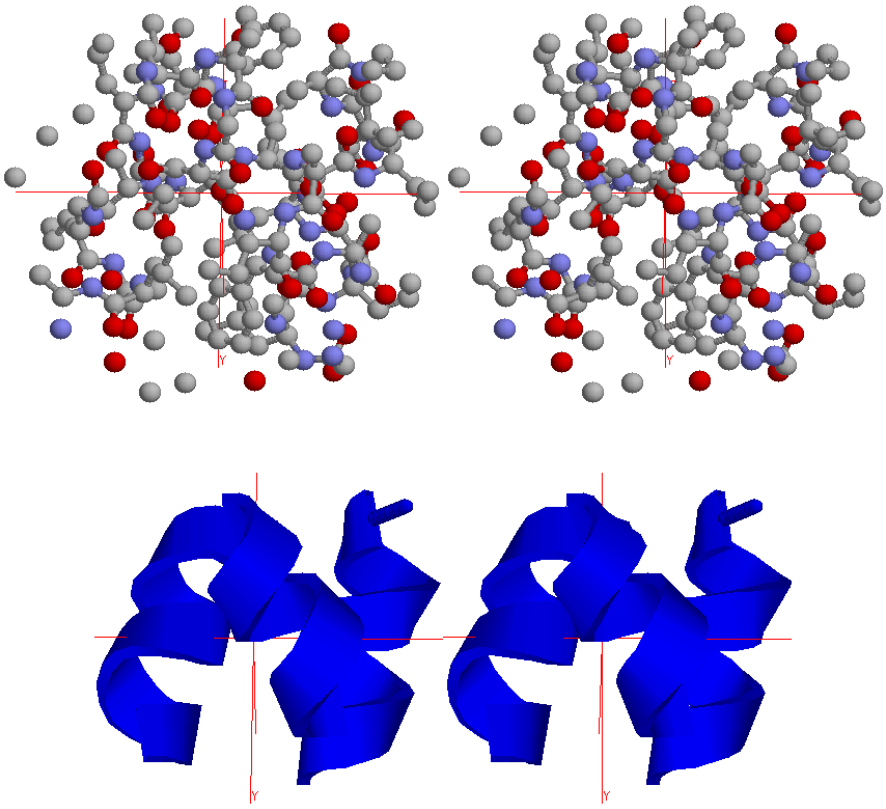

FIGURE 6.6: Stereo graph of the artificial structure displayed in ball-and-sticks and cartoons.

The space group is P43 21 2. The cell dimensions are a = 40.000Å , b = 40.000Å and c = 40.000Å . There are 25 amino acids in the asymmetric unit. The number of non-hydrogen atoms in the asymmetric unit is 229, including protein atoms, heterogen atoms and some fixed solvent atoms. The solvent content is less than 50.00%. Synthetic data has been calculated from the atomic model without any bulk solvent. The resolution of

the synthetic data is 1.2Å. The number of unique reflections is 10658. There are 5572 unique reflections above 1.5Å and 3293 unique reflections above 1.8Å. The completeness of the synthetic data is 100%, because no beam stop has been considered. F000 is automatically reconstructed during the iterations. The unit cell is discretized into a $80 \times 80 \times 80$ grid for fast Fourier transform. The distance between two nearby grid points is about 0.5Å. The density function is defined on the grid points.

In each iteration cycle, the weighted average density is calculated to construct a slightly loose molecular mask which includes 52% of the unit-cell volume. Correspondingly, the solvent region occupies 48% of the unit-cell volume. Because gradient-histogram matching has reduced the freedom of the electron density, the oversampling condition is still satisfied.

A limited hybrid input-output method has been used to modify the calculated electron density in the solvent region for all iterations. The initial value of the limited density is 0.8e/Å3 . This value linearly decreases to 0.2e/Å3 at the end of 5000 iterations. The reference gradient histogram is computed from the artificial molecule itself.

For 1.2Å synthetic data, the evolution of the mean error in phase angle is shown in Figure 6.7. All synthetic data has been used as working data set. There is no free R factor. 13 successful runs have been obtained among 1000. The success rate is low because the solvent content is low. Because synthetic data is used, the final mean error in phase angle is about 20° for 1.2 Å data.

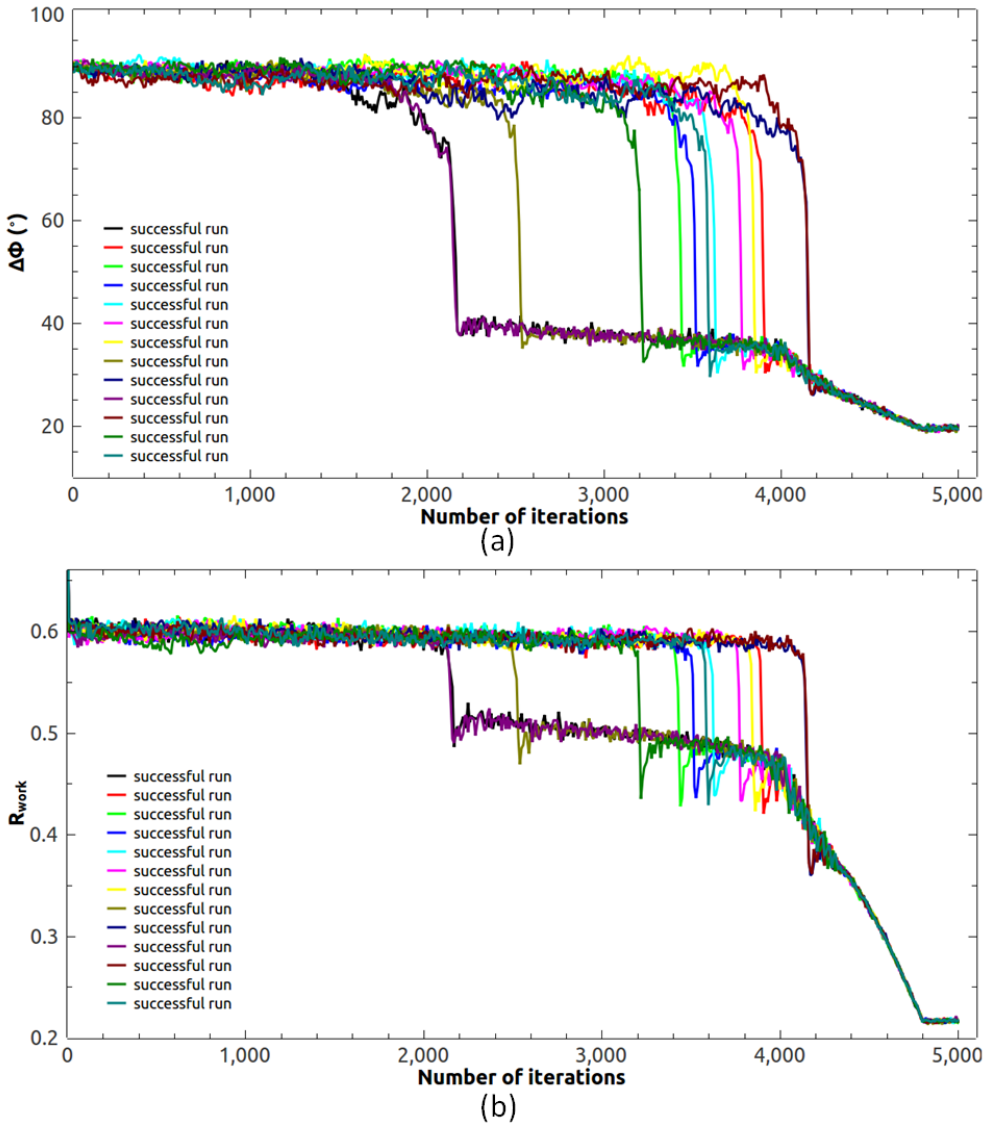

FIGURE 6.7: Evolution of the phase error and R values of an artificial protein structure using 1.2Å data.

For 1.5 Å synthetic data, a molecular mask is constructed from the weighted average density map in each iteration cycle. A limited hybrid input-output method has also been used with the same parameters. 2D histogram matching is applied. About 1% synthetic reflections are set aside to calculate the free R factor. The evolution of the R values and the mean error in phase angle has been shown in Figure 6.8. There are 3 successful

runs among 1000. The success rate becomes lower because the resolution of the data has decreased to 1.5 Å. The final mean error in phase angle is about 25° for 1.5 Å data. The calculated electron density maps are in Figure 6.9.

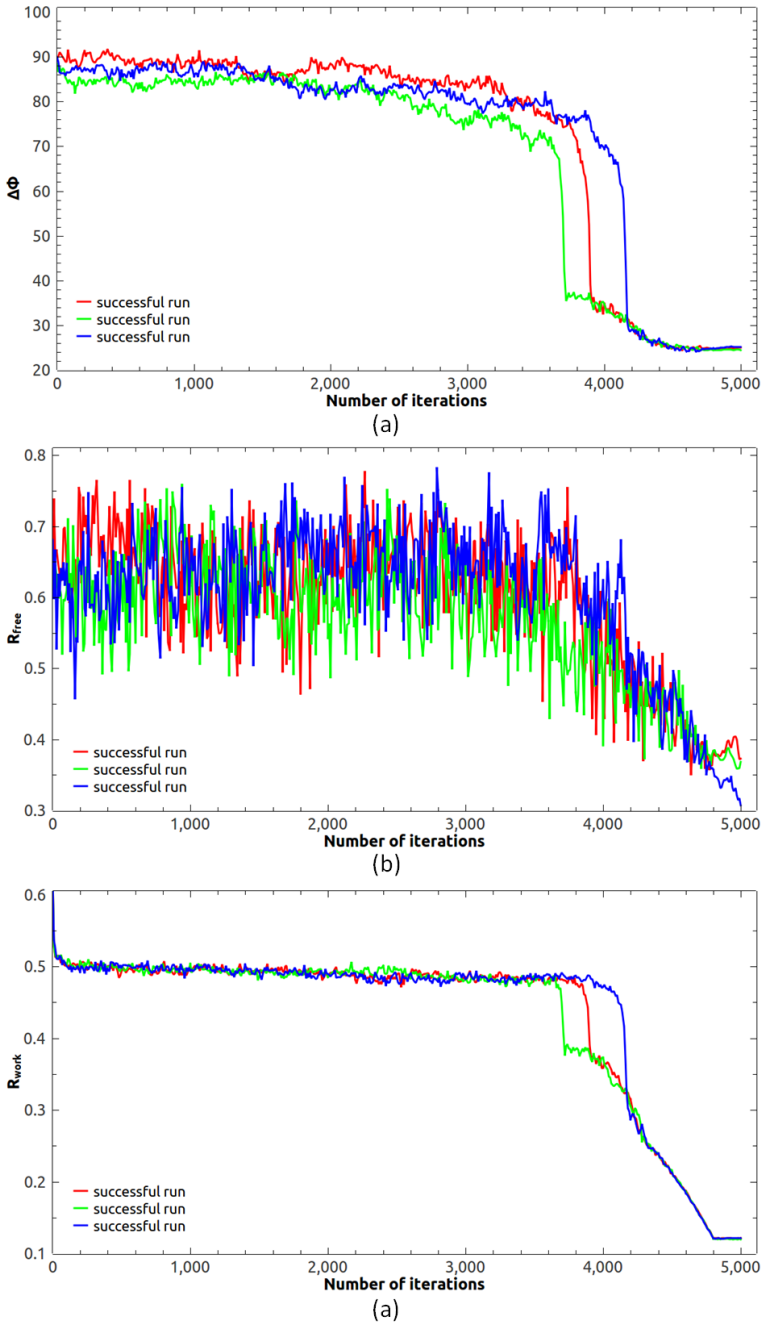

FIGURE 6.8: Evolution of the phase error and R values of an artificial protein structure using 1.5Å data.

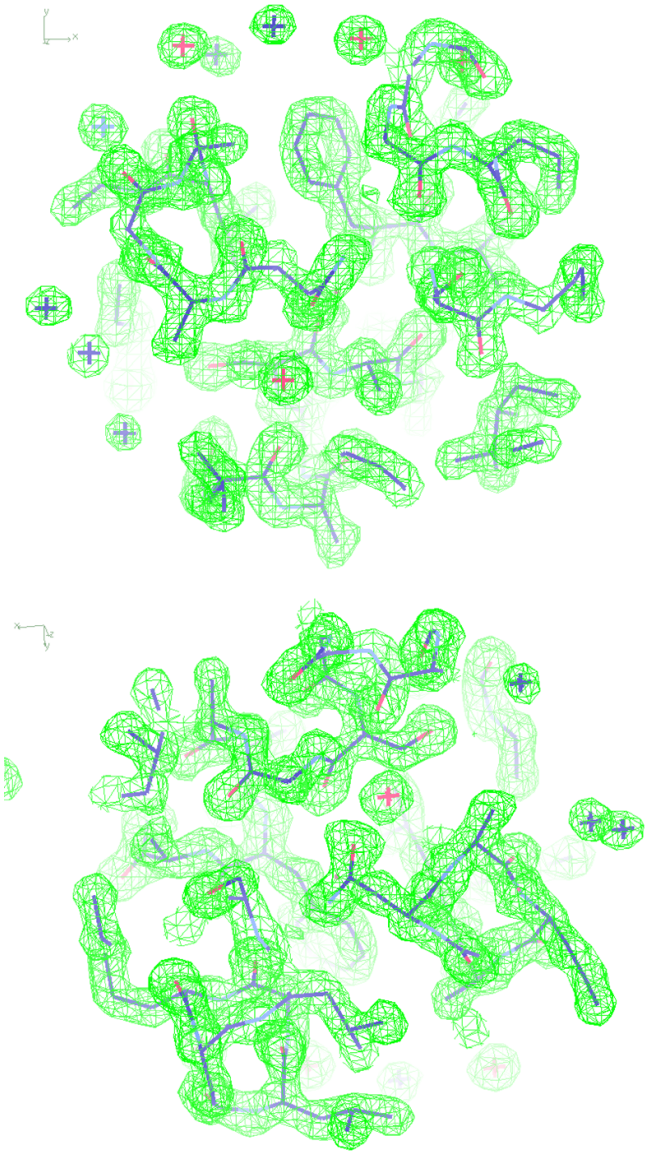

FIGURE 6.9: Calculated electron density maps (green) of 1.5Å synthetic data. The artificial structure has been superimposed (blue and red).

For 1.8 Å synthetic data, a molecular mask is also automatically constructed from the weighted average density map in each iteration cycle. A limited hybrid input-output method has also been used with the same parameters. 2D histogram matching is applied inside the molecular mask. About 1% synthetic reflections are set aside to calculate the free R factor. The evolution of the R values and the mean error in phase angle is shown

in Figure 6.10. There is only one successful run among 1000. The success rate becomes extremely low because the resolution of the data has decreased to 1.8 Å.

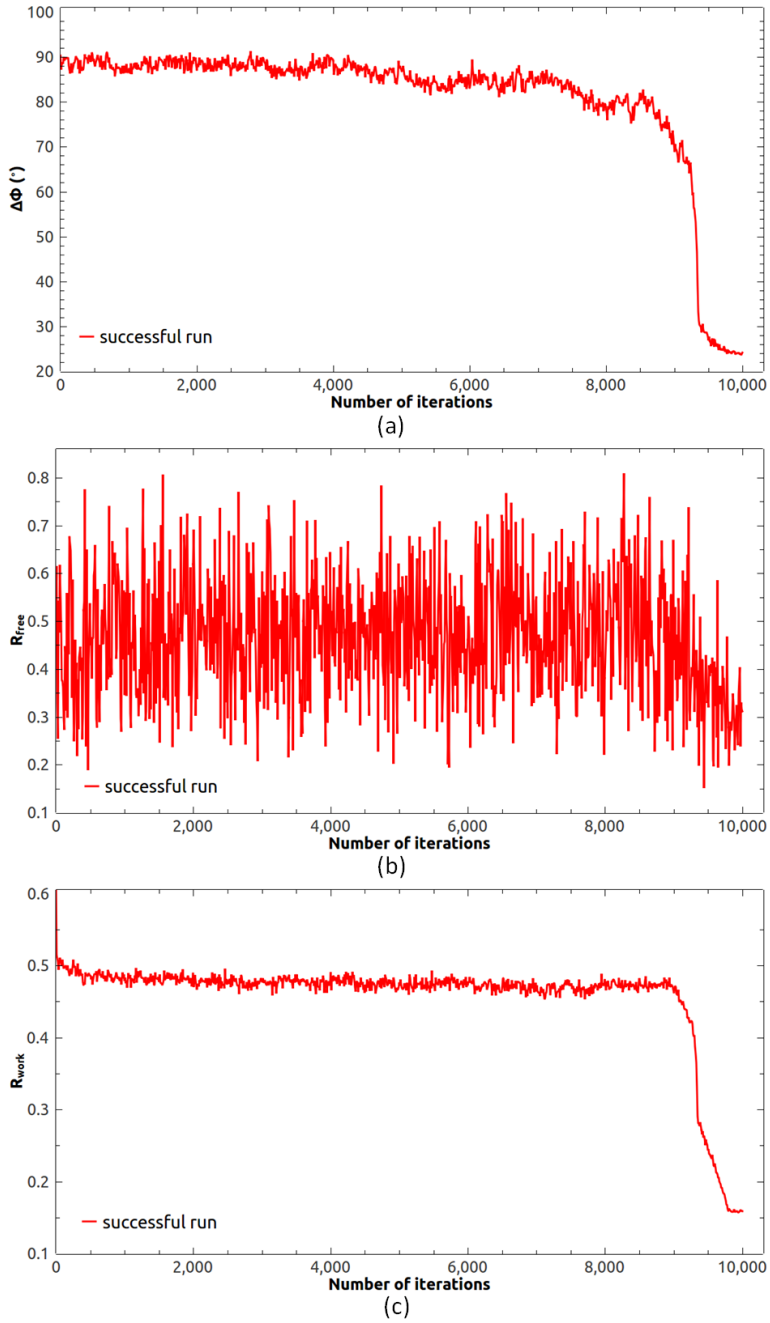

FIGURE 6.10: Evolution of the phase error and R values of an artificial structure using 1.8Å data.

The final mean error in phase angle is about 24° for 1.8Å data. The calculated electron density maps are shown in Figure 6.11.

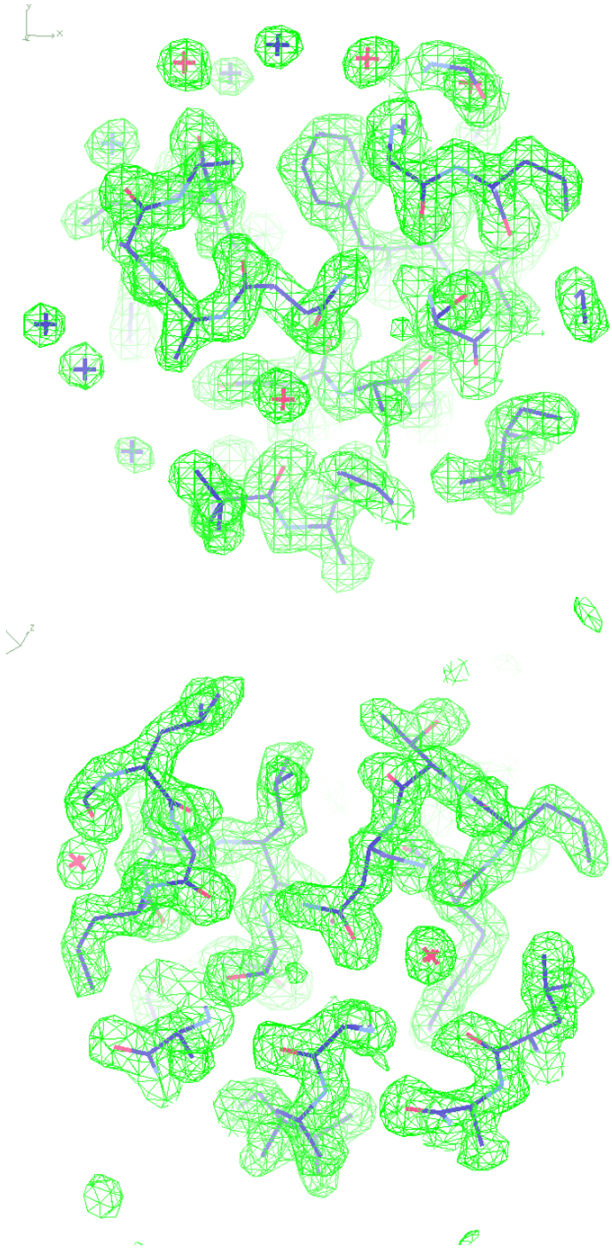

FIGURE 6.11: Calculated electron density maps (green) of 1.8Å synthetic data. The artificial structure has been superimposed (blue and red).

6.4 Conclusions

Non-crystallographic symmetry can be used modify the electron density in the asymmetric unit. NCS averaging can dramatically improve the calculated electron density. NCS has reduced the degrees of freedom of the electron density. The independent protein volume becomes much smaller. The oversampling condition can be satisfied for low-solvent-content crystals. The structure of 4NF2 with a three-fold NCS has been successfully solved directly from the synthetic data at 1.74Å resolution. The solvent content of the crystal is about 45%.

In addition to NCS, gradient constraint can also reduce the degrees of freedom of the protein electron density. The number of independent unknown variables has been reduced. The oversampling condition can be satisfied for crystals with small solvent volume. The structure of a small artificial molecule without NCS has been solved directly from the synthetic data at three resolution levels. The solvent content of the artificial molecule is less than 50%.

More constraints will be included in the future to solve protein crystals with solvent content less than 50%.

Bibliography

[1] Jan Drenth. *Principles of protein X-ray crystallography*. Springer Science & Business Media, 2007.

[2] Bernhard Rupp. *Biomolecular crystallography: principles, practice, and application to structural biology*. Garland Science, 2009.

[3] Gale Rhodes. *Crystallography made crystal clear: a guide for users of macromolecular models*. Academic press, 2010.

[4] James D Watson, Amy A Caudy, Richard M Myers, and Jan A Witkowski. *Recombinant DNA: genes and genomes: a short course*. WH Freeman New York, NY, 2007.

[5] Brunger Axel T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, 355(6359):472–475, 1992. 10.1038/355472a0.

[6] R. Diamond. A real-space refinement procedure for proteins. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 27(5):436–452, 1971.

[7] G. N. Ramachandran, C. T. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1):95–99, 1963.

[8] M. M. Woolfson. *X-ray Crystallography*. Cambridge University Press, 1970.

[9] Carmelo Giacovazzo. *Direct methods in crystallography*. Academic Press, 1980.

[10] M. S. Smyth and J. H. J Martin. X-ray crystallography. *Molecular Pathology*, 53 (1):8, 2000.

[11] Jian-Sheng Jiang and Axel T Brünger. Protein hydration observed by X-ray diffraction: solvation properties of penicillopepsin and neuraminidase crystal structures. *Journal of Molecular Biology*, 243(1):100–115, 1994.

[12] D. Waasmaier and A. Kirfel. New analytical scattering-factor functions for free atoms and ions. *Acta Crystallographica Section A: Foundations of Crystallography*, 51(3):416–431, 1995.

[13] Paul D Adams, Pavel V Afonine, Gábor Bunkóczi, Vincent B Chen, Ian W Davis, Nathaniel Echols, Jeffrey J Headd, L-W Hung, Gary J Kapral, Ralf W Grosse-Kunstleve, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D: Biological Crystallography*, 66(2):213–221, 2010.

[14] Priyamvada Acharya, Meike Goenrich, Christoph H. Hagemeier, Ulrike Demmer, Julia A. Vorholt, Rudolf K. Thauer, and Ulrich Ermler. How an enzyme binds the

C1 carrier tetrahydromethanopterin. Structure of the tetrahydromethanopterin-dependent formaldehyde-activating enzyme (Fae) from Methylobacterium extorquens AM1. *Journal of Biological Chemistry*, 280(14):13712–13719, 2005.

[15] J. T. Karle and H. Hauptman. The phases and magnitudes of the structure factors. *Acta Crystallographica*, 3(3):181–187, 1950.

[16] D. Sayre. The squaring method: a new method for phase determination. *Acta Crystallographica*, 5(1):60–65, 1952.

[17] Herbert Aaron Hauptman and Jerome Karle. *Solution of the phase problem. I. The centrosystemmetric crystal*. Number 3. American Crystallographic Association, 1953.

[18] W. T. Cochran. Relations between the phases of structure factors. *Acta Crystallographica*, 8(8):473–478, 1955.

[19] J. T. Karle and I. L. Karle. The symbolic addition procedure for phase determination for centrosymmetric and non-centrosymmetric crystals. *Acta Crystallographica*, 21(6):849–859, 1966.

[20] Peter S White and M. M. Woolfson. The application of phase relationships to complex structures. VlI. Magic integers. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 31(1):53–56, 1975.

[21] Peter Main. On the application of phase relationships to complex structures. XI. A theory of magic integers. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 33(5):750–757, 1977.

[22] Hendrik Schenk. *An Introduction to Direct Methods: The Most Important Phase Relationships and Their Application in Solving the Phase Problem*. University College Cardiff Press, 1984.

[23] Ernst Egert and George M Sheldrick. Search for a fragment of known geometry by integrated Patterson and direct methods. *Acta Crystallographica Section A: Foundations of Crystallography*, 41(3):262–268, 1985.

[24] V. Yu Lunin, A. G. Urzhumtsev, and T. P. Skovoroda. Direct low-resolution phasing from electron-density histograms in protein crystallography. *Acta Crystallographica Section A: Foundations of Crystallography*, 46(7):540–544, 1990.

[25] Axel T Brunger. Simulated annealing in crystallography. *Annual Review of Physical Chemistry*, 42(1):197–223, 1991.

[26] Charles M Weeks, George T DeTitta, Russ Miller, and Herbert A Hauptman. Application of the minimal principle to peptide structures. *Acta Crystallographica Section D: Biological Crystallography*, 49(1):179–181, 1993.

[27] A. L. Patterson. A Fourier series method for the determination of the components of interatomic distances in crystals. *Phys. Rev.*, 46:372–376, 1934.

[28] Michael G Rossmann. The molecular replacement method. *Acta Crystallographica Section A: Foundations of Crystallography*, 46(2):73–82, 1990.

[29] J. Navaza. AMoRe: an automated package for molecular replacement. *Acta Crystallographica Section A: Foundations of Crystallography*, 50(2):157–163, 1994.

[30] W.-P. Su. Simulated annealing as a tool for Ab initio phasing in X-ray crystallography. *Acta Crystallographica Section A: Foundations of Crystallography*, 51(6): 845–849, 1995.

[31] W.-P. Su. Simulated annealing and the X-ray phase problem. *Physica A: Statistical Mechanics and its Applications*, 221(1):193–201, 1995.

[32] V. Yu Lunin, N. L. Lunina, T. E. Petrova, E. A. Vernoslova, A. G. Urzhumtsev, and A. D. Podjarny. On the ab initio solution of the phase problem for macromolecules at very low resolution: the few atoms model method. *Acta Crystallographica Section D: Biological Crystallography*, 51(6):896–903, 1995.

[33] George M Sheldrick and Robert O Gould. Structure solution by iterative peaklist optimization and tangent expansion in space group P1. *Acta Crystallographica Section B: Structural Science*, 51(4):423–431, 1995.

[34] William B Drendel, Rakhal D Dave, and Sanjeev Jain. Forced coalescence phasing: a method for ab initio determination of crystallographic phases. *Proceedings of the National Academy of Sciences*, 92(2):547–551, 1995.

[35] David A Langs, Russ Miller, Herbert A Hauptman, and G. W. Han. Use of the minimal function for partial structure development in direct methods. *Acta Crystallographica Section A: Foundations of Crystallography*, 51(1):81–87, 1995.

[36] Y.-S. Chen, W.-P. Su, S. P. Mallela, and R. A. Geanangel. Solution of a new molecular structure {[SnClGe (SiC3H9) 3] 4} by simulated annealing. *Acta Crystallographica Section A: Foundations of Crystallography*, 53(3):396–399, 1997.

[37] Herbert Hauptman. Phasing methods for protein crystallography. *Current Opinion in Structural Biology*, 7(5):672–680, 1997.

[38] G. D. Smith, R. H. Blessing, S. E. Ealick, J. C. Fontecilla-Camps, H. A. Hauptman, D. Housset, D. A. Langs, and R. Miller. *Ab initio* structure determination and refinement of a scorpion protein toxin. *Acta Crystallographica Section D: Biological Crystallography*, 53(5):551–557, 1997.

[39] V. Yu Lunin, N. L. Lunina, T. E. Petrova, A. G. Urzhumtsev, and A. D. Podjarny. On the ab initio solution of the phase problem for macromolecules at very low resolution. II. Generalized likelihood based approach to cluster discrimination. *Acta Crystallographica Section D: Biological Crystallography*, 54(5):726–734, 1998.

[40] Michael F Zimmer and Wu-Pei Su. Statistical characterization of simulated annealing applied to the X-ray phase problem. *Physical Review E*, 58(4):5131, 1998.

[41] Xin Wang, Y-S Chen, and W-P Su. A real-space approach to the solution of a disordered structure of the three-dimensional coordination polymer Fe (CN) 6 [Sn (C4H9) 3] 3. *Journal of Applied Crystallography*, 32(3):409–412, 1999.

[42] Xiangan Liu and W.-P. Su. A hybrid minimal principle for the crystallographic phase problem. *Acta Crystallographica Section A: Foundations of Crystallography*, 56(6):525–528, 2000.

[43] Yan Chen and W.-P. Su. Solving the Sayre equation by simulated annealing. *Acta Crystallographica Section A: Foundations of Crystallography*, 56(2):127–131, 2000.

[44] Xiangan Liu and W.-P. Su. Improved Monte Carlo sampling in a real space approach to the crystallographic phase problem. *Phys. Rev. E*, 66:066703, Dec 2002.

[45] Yi Zhou and W.-P. Su. Solving the Sayre equations for centrosymmetric structures with a genetic algorithm. *Acta Crystallographica Section A: Foundations of Crystallography*, 60(4):306–310, 2004.

[46] W.-P. Su. Retrieving low-and medium-resolution structural features of macromolecules directly from the diffraction intensities–a real-space approach to the X-ray phase problem. *Acta Crystallographica Section A: Foundations of Crystallography*, 64(6):625–630, 2008.

[47] George M Sheldrick. A short history of SHELX. *Acta Crystallographica Section A: Foundations of Crystallography*, 64(1):112–122, 2007.

[48] Isabel Usón and George M Sheldrick. Advances in direct methods for protein crystallography. *Current Opinion in Structural Biology*, 9(5):643–648, 1999.

[49] Vladimir Y Lunin, Natalia L Lunina, Marco S Casutt, Kevin Knoops, Christiane Schaffitzel, Julia Steuber, Guenter Fritz, and Manfred W Baumstark. Low-resolution structure determination of Na+-translocating NADH: ubiquinone oxidoreductase from Vibrio cholerae by ab initio phasing and electron microscopy. *Acta Crystallographica Section D: Biological Crystallography*, 68(6):724–731, 2012.

[50] Herbert Hauptman. A minimal principle in X-ray crystallography: starting in a small way. *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences*, 442(1914):3–12, 1993.

[51] K. Y. J. Zhang and P. Main. The use of Sayre's equation with solvent flattening and histogram matching for phase extension and refinement of protein structures. *Acta Crystallographica Section A: Foundations of Crystallography*, 46(5):377–381, 1990.

[52] Russ Miller, George T DeTitta, Rob Jones, David A Langs, Charles M Weeks, and Herbert A Hauptman. On the application of the minimal principle to solve unknown structures. *Science*, 259(5100):1430–1433, 1993.

[53] George T Detitta, Charles M Weeks, Pamhla Thuman, Russ Miller, and Herbert A Hauptman. Structure solution by minimal-function phase refinement and Fourier filtering. I. Theoretical basis. *Acta Crystallographica Section A: Foundations of Crystallography*, 50(2):203–210, 1994.

[54] C. M. Weeks, Gtowji-T DeTitta, H. A. Hauptman, Pamela Thuman, and Russ Miller. Structure solution by minimal-function phase refinement and Fourier filtering. II. Implementation and applications. *Acta Crystallographica Section A: Foundations of Crystallography*, 50(2):210–220, 1994.

[55] Russ Miller, Steven M Gallo, HG Khalak, and CM Weeks. SnB: crystal structure determination via shake-and-bake. *Journal of Applied Crystallography*, 27(4):613–621, 1994.

[56] Charles M Weeks and Russ Miller. SnB: applying Shake-and-Bake to proteins. *Proceedings of the Macromolecular Crystallography Computing School*, pages 138–147, 1997.

[57] Steven E Ealick. Now we're cooking: new successes for shake-and-bake. *Structure*, 5(4):469–472, 1997.

[58] Ashley M Deacon, Charles M Weeks, Russ Miller, and Steven E Ealick. The Shake-and-Bake structure determination of triclinic lysozyme. *Proceedings of the National Academy of Sciences*, 95(16):9284–9289, 1998.

[59] Charles M Weeks and Russ Miller. Optimizing Shake-and-Bake for proteins. *Acta Crystallographica Section D: Biological Crystallography*, 55(2):492–500, 1999.

[60] S. Subbiah. Low-resolution real-space envelopes: an approach to the ab initio macromolecular phase problem. *Science*, 252(5002):128–133, 1991.

[61] Robert D Oeffner, Gábor Bunkóczi, Airlie J McCoy, and Randy J Read. Improved estimates of coordinate error for molecular replacement. *Acta Crystallographica Section D: Biological Crystallography*, 69(11):2209–2215, 2013.

[62] Pavel Strop, Michael R Brzustowicz, and Axel T Brunger. Ab initio molecular-replacement phasing for symmetric helical membrane proteins. *Acta Crystallographica Section D: Biological Crystallography*, 63(2):188–196, 2007.

[63] Z-C Liu, Rui Xu, and Y-H Dong. Phase retrieval in protein crystallography. *Acta Crystallographica Section A: Foundations of Crystallography*, 68(2):256–265, 2012.

[64] James R Fienup. Phase retrieval algorithms: a comparison. *Applied optics*, 21 (15):2758–2769, 1982.

[65] Hongxing He and Wu-Pei Su. Direct phasing of protein crystals with high solvent content. *Acta Crystallographica Section A: Foundations and Advances*, 71(1):92–98, 2015.

[66] R. P. Millane and W. J. Stroud. Reconstructing symmetric images from their undersampled Fourier intensities. *JOSA A*, 14(3):568–579, 1997.

[67] J. L. van der Plas and Rick P Millane. Ab-initio phasing in protein crystallography. In *International Symposium on Optical Science and Technology*, pages 249–260. International Society for Optics and Photonics, 2000.

[68] Stefano Marchesini, H. He, Henry N. Chapman, Stefan P. Hau-Riege, A. Noy, Malcolm R. Howells, U. Weierstall, and John C. H. Spence. X-ray image reconstruction from a diffraction pattern alone. *Physical Review B*, 68(14):140101, 2003.

[69] Roman Dronyak, Keng S Liang, Yuri P Stetsko, Ting-Kuo Lee, Chi-Kai Feng, Jin-Sheng Tsai, and Fu-Rong Chen. Electron diffractive imaging of nano-objects using a guided method with a dynamic support. *Applied Physics Letters*, 95(11): 111908, 2009.

[70] J. Miao, D. Sayre, and H. N. Chapman. Phase retrieval from the magnitude of the Fourier transforms of nonperiodic objects. *JOSA A*, 15(6):1662–1669, 1998.

[71] Theo Hahn, Uri Shmueli, Arthur James Cochran Wilson, and Edward Prince. *International Tables for Crystallography*. D. Reidel Publishing Company, 2005.

[72] Lucas Bleicher, Ricardo Aparicio, Fabio M Nunes, Leandro Martinez, Sandra M Gomes Dias, Ana CM Figueira, Maria AM Santos, Walter H Venturelli, Rosangela Da Silva, Paulo M Donate, et al. Structural basis of GC-1 selectivity for thyroid hormone receptor isoforms. *BMC Structural Biology*, 8(1):8, 2008.

[73] Juergen Koepke, Eva-Maria Krammer, Astrid R Klingen, Pierre Sebban, G Matthias Ullmann, and Günter Fritzsch. pH modulates the quinone position in the photosynthetic reaction center from Rhodobacter sphaeroides in the neutral and charge separated states. *Journal of Molecular Biology*, 371(2):396–409, 2007.

[74] Natalia Lunina, V Lunin, and Alexandre Urzhumtsev. Connectivity-based ab initio phasing: from low resolution to a secondary structure. *Acta Crystallographica Section D: Biological Crystallography*, 59(10):1702–1715, 2003.

[75] Andrei Fokine, Natalia Lunina, Vladimir Lunin, and Alexandre Urzhumtsev. Connectivity-based ab initio phasing at different solvent levels. *Acta Crystallographica Section D: Biological Crystallography*, 59(5):850–858, 2003.

[76] K. Y. J. Zhang and Peter Main. Histogram matching as a new density modification technique for phase refinement and extension of protein molecules. *Acta Crystallographica Section A: Foundations of Crystallography*, 46(1):41–46, 1990.

[77] K. Y. J. Zhang and P. Main. The use of Sayre's equation with solvent flattening and histogram matching for phase extension and refinement of protein structures. *Acta Crystallographica Section A: Foundations of Crystallography*, 46(5):377–381, 1990.

[78] Stefano Marchesini. Invited article: A unified evaluation of iterative projection algorithms for phase retrieval. *Review of Scientific Instruments*, 78(1):011301, 2007.

[79] Bi-Cheng Wang. Resolution of phase ambiguity in macromolecular crystallography. *Methods in Enzymology*, 115:90–112, 1985.

[80] Thomas C Terwilliger. Reciprocal-space solvent flattening. *Acta Crystallographica Section D: Biological Crystallography*, 55(11):1863–1871, 1999.

[81] J. P. Abrahams. Bias reduction in phase refinement by modified interference functions: introducing the γ correction. *Acta Crystallographica Section D: Biological Crystallography*, 53(4):371–376, 1997.

[82] Kevin D. Cowtan and Kam Y. J. Zhang. Density modification for macromolecular phase improvement. *Progress in Biophysics and Molecular Biology*, 72(3):245–270, 1999.

[83] Susana LA Andrade, Eric V Patridge, James G Ferry, and Oliver Einsle. Crystal structure of the NADH: quinone oxidoreductase WrbA from *Escherichia coli*. *Journal of Bacteriology*, 189(24):9101–9107, 2007.

[84] Minxie Qian, El Hassan Ajandouz, Françoise Payan, and Virginie Nahoum. Molecular basis of the effects of chloride ion on the acid-base catalyst in the mechanism of pancreatic α-amylase. *Biochemistry*, 44(9):3194–3201, 2005.

[85] Winfried Meining, Simone Mörtl, Markus Fischer, Mark Cushman, Adelbert Bacher, and Rudolf Ladenstein. The atomic structure of pentameric lumazine

synthase from Saccharomyces cerevisiae at 1.85 Å resolution reveals the binding mode of a phosphonate intermediate analogue. *Journal of Molecular Biology*, 299 (1):181–197, 2000.

[86] Victor Lo, Richard L Kingston, and RP Millane. Determination of molecular envelopes from solvent contrast variation data. *Acta Crystallographica Section A: Foundations of Crystallography*, 65(4):312–318, 2009.

[87] Jianwei Miao, Pambos Charalambous, Janos Kirz, and David Sayre. Extending the methodology of X-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. *Nature*, 400(6742):342–344, 1999.

[88] Jianwei Miao, J Kirz, and D Sayre. The oversampling phasing method. *Acta Crystallographica Section D: Biological Crystallography*, 56(10):1312–1315, 2000.

[89] Jianwei Miao and D Sayre. On possible extensions of X-ray crystallography through diffraction-pattern oversampling. *Acta Crystallographica Section A: Foundations of Crystallography*, 56(6):596–605, 2000.

[90] Chien-Chun Chen, Jianwei Miao, C. W. Wang, and T. K. Lee. Application of optimization technique to noncrystalline x-ray diffraction microscopy: Guided hybrid input-output method. *Phys. Rev. B*, 76:064113, Aug 2007.

[91] Michael G Rossmann. Ab initio phase determination and phase extension using non-crystallographic symmetry. *Current Opinion in Structural Biology*, 5(5):650–655, 1995.

[92] Garry L Taylor. Introduction to phasing. *Acta Crystallographica Section D: Biological Crystallography*, 66(4):325–338, 2010.

[93] Rick P Millane and Victor L Lo. Iterative projection algorithms in protein crystallography. I. Theory. *Acta Crystallographica Section A: Foundations of Crystallography*, 69(5):517–527, 2013.

[94] Veit Elser. Solution of the crystallographic phase problem by iterated projections. *Acta Crystallographica Section A*, 59(3):201–209, May 2003.

[95] Veit Elser. Phase retrieval by iterated projections. *JOSA A*, 20(1):40–55, 2003.

[96] Quan Hao. Phasing from an envelope. *Acta Crystallographica Section D: Biological Crystallography*, 57(10):1410–1414, 2001.

[97] Brian W Matthews. Solvent content of protein crystals. *Journal of Molecular Biology*, 33(2):491–497, 1968.

[98] Katherine A Kantardjieff and Bernhard Rupp. Matthews coefficient probabilities: improved estimates for unit cell contents of proteins, DNA, and protein–nucleic acid complex crystals. *Protein Science*, 12(9):1865–1871, 2003.

[99] Alexandra Goldstein and Kam YJ Zhang. The two-dimensional histogram as a constraint for protein phase Improvement. *Acta Crystallographica Section D: Biological Crystallography*, 54(6):1230–1244, 1998.

[100] Y.-P. Nieh and Kam Y. J. Zhang. A two-dimensional histogram-matching method for protein phase refinement and extension. *Acta Crystallographica Section D: Biological Crystallography*, 55(11):1893–1900, 1999.