Structure of Bovine Liver Catalase Solved by Electron Diffraction on Multilayered Crystals

A Dissertation

Submitted to the Department of Physics
in Partial Fulfilment of the Requirements for the
Degree of *Doctor Rerum Naturalium*

by

Victor Kulik

UNIVERSITÄT OSNABRÜCK

May, 2004

Structure of Bovine Liver Catalase Solved
by Electron Diffraction on Multilayered Crystals

The high resolution structure of protein molecules and protein-protein complexes is important to investigate their functions. Today, large 3D or 2D crystals are required to obtain protein structures by X-ray crystallography or conventional Electron Microscopy, respectively. However, production of such crystals of good quality is a solely empirical procedure, which relies on screening numerous crystallization conditions. At the same time, multilayered protein crystals are often a by-product of attempts to grow 3D or 2D crystals and could be obtained more easily. So far, multilayered protein crystals have not been used in electron microscopy for structure determination, as the interpretation of an electron diffraction pattern is rather complicated.

In this thesis we present the first protein structure – bovine liver catalase at 4 Å resolution – solved using electron diffraction data from multilayered crystals. 55 diffraction patterns (17 tilt series) were recorded and used for the reconstruction. The tilt geometry of each individual diffraction pattern was determined by a least-squares algorithm or Laue zone analysis to perform spot indexing. The phase problem was solved by molecular replacement. The influence of the missing data cone on the self-rotation function and interpretation of reconstructed map is discussed.

dedicated to

Grigorij Borisenko

(1916-1993)

Contents

List of Figures

List of Abbreviations:

2D	two-dimensional
3D	three-dimensional
CC	correlation coefficient
CCD	couple charge device
CCP4	Collaborative Computing Project Number 4
CNS	Crystallograpgy & NMR system
CPU	central processing unit
DNA	deoxyribose aucleic acid
EC	electron crystallography
EELS	electron energy loss spectroscopy
EFTEM	energy-filtered transmission electron microscope
EM	electron microscopy
FEL	free-electron laser
FFT	Fast Fourier transform
IP	Imaging Plate
MAD	multiwavelength anomalous dispersion
MIR	multiple isomorphous replacement
MR	molecular replacement
MW	molecular weight
NCS	non-crystallographic symmetry
NMR	nuclear magnetic resonance
PDB	Protein Data Bank
PEG	polyethylene glycol
RNA	ribonucleic acid
S/N	signal-to-noise
SAD	single anomalous dispersion
SIRAS	single isomorphous replacement and anomalous scattering
SLCC	standard linear correlation coefficient
TEM	transmission electron microscope

Chapter 1

Introduction

"There is nothing over which a free man ponders less than death; his wisdom is to mediate not on death but on life" (*1*). Erwin Schrödinger chose these Spinoza words as an epigraph to the book "What is life? The physical aspect of the living cell" (*2*). Actually, what is life? Scientists, from theologists and philosophers to biologists and physicists, have tried to answer this question for many centuries. Although information about living systems increases dramatically, the understanding of life still remains unclear.

To answer Schrödinger's question, we have to address several questions: What is the nature of processes in the living system? What is the structure of a living system and its pattern of organization? Nowadays, it becomes clear that all "primary" processes in the living system are governed by the same laws of chemistry and physics as in the abiocoen[1]. Indeed, all living organisms consist of atoms – or better – molecules, which obey physical laws.

However, if we compare chemical reactions in solution and in a cell, we will find that the reactions in the cell often follow a completely different kinetic pathway. One of the most spectacular examples is that of decarboxylation of orotic acid[2]. It spontaneously decarboxylates with $t_{1/2}$ of 78 million years at room temperature in neutral aqueous solution. In the living organism the same reaction takes place faster by 10^{17} fold (*3*). The general reason for such acceleration is a hierarchy of structural organization reaching from the 3D molecular structure of biocatalytic molecules to the compartmentalization of cells.

---

[1] Non-living matter.

[2] A structural formula of orotic acid:
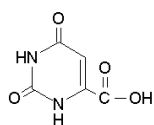
1.1 Proteins as Primary Building Blocks

What are the building blocks of the molecular machinery in a cell? This role is carried out by special macromolecules called proteins. A typical cell may contain thousands of different proteins at any time, which perform a variety of tasks in the cell. A large class of proteins, called enzymes, plays an essential role in catalyzing biochemical reactions in the cell. Enzymes can help to produce complex biological molecules or to break down such molecules. For example, orotidine 5′-phosphate decarboxylase catalyses the reaction described above. Other proteins, such as the cytoskeletal protein actin, serve as scaffolding in the cell and give cells shape, or help to form compartments in which different cellular functions are partitioned. Some proteins bind to nucleic acids and other cellular constituents. Special proteins also function as hormones (for example, insulin) or antibodies, and some are involved in transportation of other molecules (for example, hemoglobin, which stores oxygen in red blood cells).

This wide variety of proteins is produced by simple combinations of only 20 common building blocks, the amino acids. Amino acids are connected to each other in a linear fashion by peptide bonds. The linear sequence of amino acids is called "primary structure" of a protein. Information about the "primary structure" is encoded in the DNA sequence. The sequencing of the human genome (Human Genome Project) has been hailed as an epochal triumph of the 20th century that opens yet unbounded opportunities for advances in human healthcare (*4*). However, it is generally accepted that such progress can only occur when the genomic information is complemented by our understanding of protein function.

A complete description of a protein and its function will almost always require knowledge of its 3D structure. However, the rate at which genomic sequences are obtained has by far outstripped that of structure determination. At the present time, biologists are faced with the more formidable task to determine the structure and function of about 100.000 human gene products, *i.e.* encoded proteins.

Once the primary, linear chain of amino acids is formed, amino acids arrange to create the "secondary structure". The term "protein secondary structure" refers to common repeating structures found in proteins. There are two basic components of "secondary" structure:

alpha helices and beta-pleated sheets. "Tertiary structure" describes the next higher level of folding in which the helices and sheets of the "secondary structure" fold upon themselves forming the three-dimensional structure of the polypeptide chain. This higher level folding arises from several reasons. Firstly, different regions of the amino acid chain are hydrophilic or hydrophobic and arrange themselves accordingly in water. Secondly, different regions of the chain interact with each other via hydrogen bonding or disulfide linkages. Disulfide bond patterns as well as ionic and hydrophobic interactions greatly influence "tertiary structure".

The highest level of structural organization is the spatial arrangement of individually folded polypeptide chains, the "quaternary structure". The complex structure forms when polypeptide chains are bound together, usually by hydrogen bonds. For example hemoglobin, the oxygen carrying protein in blood, has four chains held together by noncovalent interactions forming a tetramer.

The amazing feature of protein nature is that the "primary structure" determines all higher levels of protein structure organization. The linear polypeptide chain folds in a particular arrangement, giving a three-dimensional structure. All the information about the native fold appears to be included in the "primary structure" with the consequence that a large number of proteins is self-folding (although *in vivo*, polypeptide folding is often assisted by additional molecules known as molecular chaperones).

Structures of individual macromolecules are often uninstructive about their function, if taken out of context. It is obvious that the knowledge of a single high resolution structure of an isolated protein will not be sufficient to understand its function in the cellular environment. "Just as words must be assembled into sentences, paragraphs, chapters and books to make sense, all cellular components must be organized in structural ensembles to perform vital cellular functions" (*5*). Most proteins do not act independently or statically in living systems. In carrying out their functions inside cells, proteins form complexes with other proteins. Thus, complex formation is a crucial mechanism for a cellular function. However, the role of protein structure for protein interactions and the formation of protein complexes in a cell is still poorly understood. A major research goal is not only to analyse structures of individual proteins, but rather to study the structure of protein-protein

complexes. By doing so, one hopes to understand how protein-protein interactions regulate intra- and inter-cell signalling and how the connecting networks between proteins in genomes could work.

One can suggest that the 3D structure of a single protein can be calculated theoretically if a "primary structure" is known. Unfortunately, at the present time, theoretical methods are only able to predict with high probability a localization of "secondary structure" components. Currently, the "tertiary structure" cannot be obtained solely by theoretical methods. Experimental information must be used.

Nowadays, the three following methods are well established for structure determination of macromolecules: Nuclear Magnetic Resonance (NMR), X-ray crystallography and Electron Microscopy (EM) (Chapter 1.2). However, each method has limitations for the determination of protein complex structures. Routinely, NMR can deal only with peptides and single small proteins due to protein size limitation. X-ray crystallography can provide high resolution structural information, but it needs relatively large, well-ordered protein crystals which are not always available. EM can work with single protein complexes, but it provides only subatomic resolution. This consideration leads to the question: Is it possible to develop a method for protein and protein complex structure determination which would combine advantages of X-ray analysis (high-resolution) and EM (small sample size)? Our results presented in this thesis show that Electron Crystallography of multilayered crystals could be such method.

1.2 Modern Methods to Determine Protein Structure

As biological phenomena are fundamentally defined at the molecular level, it is important to understand the structure and dynamics of biomolecules and their complexes. Since the dynamic changes such as domain movements, loop closure and opening, or changes in orientation of side chains of amino acid residues are within a few Å, it is required to determine the protein structures at quasi-atomic resolution.

So far only three methods have been used to obtain protein or protein complex structures at such resolution: NMR spectroscopy, X-ray crystallography, and Electron Crystallography (EC) of 2D crystals.

1.2.1 Nuclear Magnetic Resonance Spectroscopy

NMR spectroscopy can solve protein and peptide structures in solution, thus their structure and dynamics can be investigated under physiological conditions. About 14% of the structures in the database of biomolecular structures have been solved using this technique (*6*). NMR spectroscopy can also be used for drug screening and discovery (*6*), diagnosis of human diseases through analysis of biological fluids (*7*), and identification of active components in complex natural products (*8*). The use of NMR spectroscopy, however, is limited by the molecular weight of the proteins that can be studied. Large protein molecules usually face the problem of serious resonance overlap and loss of signal intensity due to fast relaxation. However, this size limitation decreases due to recent developments such as complex multidimensional pulse sequences, incorporation of NMR-active nuclei such as ^{15}N and ^{13}C into the molecule of interest and using magnets to allow GHz excitation.

1.2.2 X-ray Crystallography

X-ray crystallography is the oldest method to obtain macromolecular structures at atomic resolution. Currently, it is the most prolific technique for the structural analysis of proteins and protein complexes, and it still is a "gold" standard in terms of accuracy (*5*). X-ray crystallography can be applied to a wide range of structural complexity. The size of the structures determined by the method range from small proteins (~10 kDa) to the 70S ribosome, which consists of 52 proteins and 3 RNA molecules and has a relative molecular mass of ~2.500 kDa. Typically, for crystallographic analysis the protein is over-expressed

in a suitable system and purified to homogeneity. It is crystallized and diffraction data are collected using a rotating anode X-ray generator or dedicated synchrotron beamlines.

The electron density maps of the crystal are obtained using established crystallographic methods such as Molecular Replacement (MR), Multiple Isomorphous Replacement (MIR), or Multiwavelength Anomalous Dispersion (MAD). The three-dimensional protein model is then built according to the electron density map.

Although a structure contains plenty information in itself, a single, isolated structure is not enough to tell us how the molecule folds and works. This is due to the static nature of information obtained or as Frauenfelder has formulated, "… it is difficult to assess from a sitting skier how well he skis…. " (*9*). The study of reaction intermediates by X-ray diffraction, known as time-resolved crystallography, has added time as fourth dimension to protein analysis (*10-12*). The biochemical reaction is initiated in a crystal using different triggering methods and intermediate steps of the reaction can be recorded as a function of time. This approach allows to study the relationship between structure and function of proteins directly.

Unfortunately, generally, the rate-limiting step of the structural characterization of proteins or protein-protein complexes by X-ray crystallography is the crystallization procedure. High quality crystals must be produced. If large, well diffracting crystals are available, structural information can be extracted from the X-ray diffraction data. Very often the quality of crystals is an insurmountable obstacle. The larger the protein or protein-protein complex, the more difficult it can be to grow suitable crystals. Very often crystallization experiments result in the production of only small crystals (up to 10 μm in one dimension). In general, nowadays, such microcrystals are too small for the X-ray analysis and cannot be utilized in a conventional X-ray diffraction experiment, even when using microfocusing techniques and synchrotron radiation. To use these crystals, the third method for determination of a protein structure, Electron Crystallography (EC), must be applied.

1.2.3 Electron Crystallography

The electron microscope is an imaging device which exceeds the resolution power of a light microscope by several orders of magnitude due to a much shorter wavelength, which is usually in the order of a few pico-meters. Today, electron microscopy is widely used for studying the three-dimensional organization of protein and protein-protein complexes. The electron microscope produces images that represents the 2D projection of the individual molecules, the information about the third dimension is lost. Nevertheless, using special techniques, collectively called "single particle image processing", the full 3D structure of the object can be reconstructed. Although the resolving power of a modern microscope can be as good as 0.1-0.2 nm, the resolution of the reconstructed biological objects is typically an order of magnitude worse. The reasons are problems of specimen preparation, low image contrast, and radiation damage of the specimen. To decrease beam damage, the specimen's image is recorded with as low dose as possible. But this results in a low signal-noise (S/N) ratio and decreases the resolution of the reconstructed electron density map.

On the other hand, the S/N ratio is dramatically increased when a crystalline specimen is used. Since all molecules in the crystal lattice have at least the periodicity of the lattice, the signal from many identically oriented molecules is averaged. This results in a largely improved signal-to-noise-ratio (which is proportional to a square root of the number of particles averaged) providing higher resolution in the resulting reconstruction.

Protein microcrystals, which can be used for Electron Crystallography, can be divided into two main categories:

1) Two dimensional (2D) crystals which are single layers of molecules (complexes) ordered within the x,y-plane.
2) Multilayered crystals, which are several layers of unit cells thick and often a by-product of attempts to grow either 3D or 2D crystals.

2D crystals. So far, 2D crystals have been the most successful sample for obtaining high-resolution structural information by EM (*13-16*). EM of 2D crystals has been applied primarily to membrane proteins, for which the native membrane environment is a natural

"scaffold" for forming planar sheets. Images of differently tilted crystals are needed for reconstruction. Furthermore, imaging and diffraction techniques can be combined to determine a protein structure from 2D crystals.

However, working with 2D crystals has two disadvantages. Firstly, one needs to grow a perfectly planar crystal. It is not a trivial task since special crystallization conditions must be found and the crystal has to retain good quality after transfer onto the specimen holder. Secondly, series of tilted images of the crystals are required. Since each image provides only a single projected view of the molecule, one needs pictures of highly tilted crystals in order to be able to reconstruct a structure without artefacts. However, practically, the maximal tilt angle in modern microscopes is limited to $\pm$ 60° due to hardware constrains. It means, that part of the information, known as missing wedge, cannot be collected which leads to loss of resolution in the direction of the incident electron beam. This produces quite a number of difficulties for reconstructed map interpretation and model building.

Multilayered crystals. Multilayered protein crystals are often a by-product of crystallization experiments and they are an intermediate step between mono-layered (i.e. 2D) and true 3D crystals. These crystals, which are several unit cells thick, have quasi-infinite periodicity in the x,y directions and limited periodicity in the third direction. Nowadays, it is commonly accepted that thin multilayered crystals are useless for electron diffraction experiments and that only 2D crystals can be utilized by electron crystallography. Although diffraction patterns from multilayered crystals have been recorded (*17-20*), their quality was not high enough for quantitative interpretation. Several studies were initiated to solve a protein structure from multilayered crystals, but no protein model has been reported. The possible obstacles were beam damage of the sample, difficulty to collect a dataset with high completeness, and the multiple scattering effect due to the high cross-section of electrons. However, the theoretical work by Vossen (*21*) revealed that the multiple scattering effect can be corrected if the crystal thickness is less than typically 500 nm. After the symmetrization of the diffraction patterns, diffraction intensities can be used for density reconstruction and protein model building. Hence, real experiments have to prove that multilayered protein crystals can be used for structure determination.

The main advantage of Electron Crystallography over X-ray crystallography is the fact that much smaller crystals can be used for a diffraction experiment. This follows from the stronger interaction with matter of electrons than X-rays. It means that interaction of electrons even with 2D crystals produces an interpretable signal. A much stronger signal can be obtained when a multilayered crystal is used.

As it is clear from discussion above, that Electron Crystallography of multilayered protein crystals would be an ideal method for structure determination of protein complexes which may open new opportunities in structural biology. It would present such advantages as relatively simple crystal production, small crystal size and high resolution of the obtained protein structures. Since theoretical investigations showed that information about a protein structure can be extracted from electron diffraction pattern (*21-24*), the goal of this thesis is an experimental proof of principle of Electron Crystallography on multilayered protein crystals to obtain high resolution structures of biological macromolecules and protein complexes.

1.3 Scope of This Thesis

The main aim of this thesis was to experimentally prove the feasibility of electron crystallography using multilayered protein crystals as a new structure determination method. Chapter 2 focuses on physical aspects of the interactions of X-rays and electrons with matter. Chapter 3 describes an overview of modern crystallization techniques. In chapter 4, the strategy and results of collection of a 3D dataset from multilayered crystals of bovine liver catalase are presented. Chapter 5 provides the theoretical background for solving the phase problem using the Molecular Replacement method. Chapter 6 is devoted to the solution of experimental phase problem of the catalase data by MR. In particular, the effect of missing cone[3] on the phase determination and final reconstruction is presented Finally, the results and implications of the development of the new method are discussed (Chapter 7).

---

[3] As will be discussed in Chapter 4, a maximal data which can be obtained from multilayered crystals is systematically absent in one direction due to the restricted tilt geometry of the modern cryoholders. If one single crystal is used, the absent data form the "missing wedge" in the Fourier space. However, when data from many differently oriented crystals are merged, the missing data diminish to a "missing cone" in the Fourier space.

Chapter 2

Interaction with Matter

Knowing the structures of (macro)molecules at high resolution is required to understand their mechanisms. This information can be obtained by scattering experiments. In this chapter we describe the general mechanisms of the interaction of electrons and X-ray photons with matter and the theory of the scattering process. In addition, advantages and disadvantages of the use of electrons and photons for studies on biological objects are discussed.

2.1 The General Concept in a Quantum Mechanical Description: Scattering Cross-section

Let us consider the scattering of a beam of monoenergetic particles, e.g. electrons, photons or neutrons[4], at a scattering target. It is known that in the quantum mechanical description of matter the beam of incident particles can be treated as a wave (Fig. 2.1).
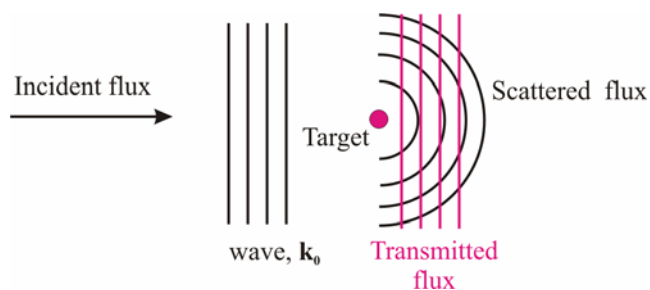
Figure 2.1: Wave mechanical description of scattering

---

[4] In the following description we follow the notation of Reimer (*25*) and Blundell and Johnson (*26*).

At far distances from the scattering center, the incoming beam is described as a plane wave:

$$\Psi_{in} = \Psi_0 e^{ik_o z} \qquad (2.1)$$

where $\mathbf{k_0}$ is the wave vector.

The incoming particle flux is then:

$$\Phi_{in} = v |\Psi_{in}|^2 \qquad (2.2)$$

$$\text{with } v = \frac{\hbar k}{m} \qquad (2.3)$$

where v is the beam velocity and m is mass of the particle.

The scattered wave at far distances from the scattering center can be described as a spherical wave:

$$\Psi_{sc} = \Psi_0 \frac{f(\theta)}{r} e^{ikr} \qquad (2.4)$$

where $f(\theta)$ is the structure factor of the target which depends on the scattering angle θ.

The corresponding flux can be expressed as:

$$\Phi_{sc} = v |\Psi_{sc}|^2 \qquad (2.5)$$
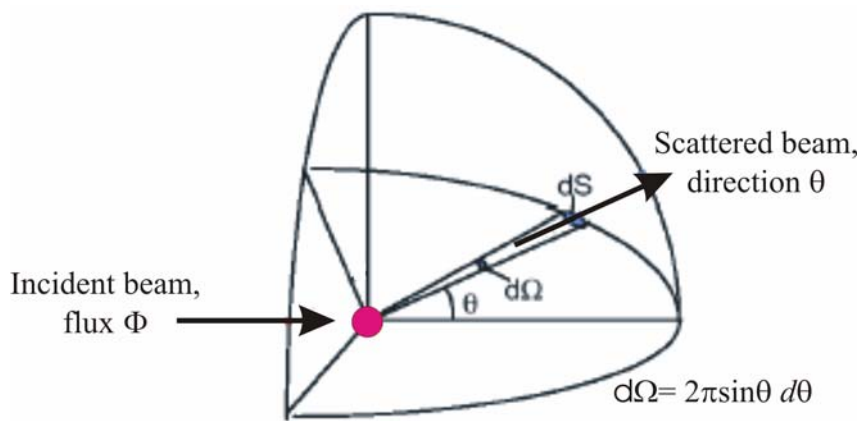
Figure 2.2: Scattering of a beam of particles through the surface element dS.

Next consider the flux $\widetilde{\Phi}$ of particles through the surface element $dS = r^2 d\Omega$ (Fig. 2.2):

$$\tilde{\Phi} = \Phi_{sc}dS = v\left|f(\theta)\frac{e^{ikr}}{r}\right|^2 dS = v|f(\theta)|^2\frac{dS}{r^2} = v|f(\theta)|^2 d\Omega \qquad (2.6)$$

The differential cross-section is defined as the number of particles scattered into $d\Omega$ per unit time and flux:

$$\frac{d\sigma}{d\Omega} = \frac{d}{d\Omega}\left(\frac{\tilde{\Phi}}{\Phi}\right) = \frac{d}{d\Omega}\left(\frac{v|f(\theta)|^2 d\Omega}{v}\right) = |f(\theta)|^2 \qquad (2.7)$$

In consequence the total scattering cross-section is given by

$$\sigma_{tot} = \int_0^{4\pi}\frac{d\sigma}{d\Omega}d\Omega \qquad (2.8)$$

2.2 Atomic and Molecular Structure Factors

Let us consider the coherent scattering of the incident wave by an atom. In order to obtain the expression for scattering by an atom it is necessary to take into account the spatial distribution of the scatterers, electrons and nucleus. This spatial distribution is described by the potential V($\mathbf{r}$).

Let the density (electron or nucleus density) at a distance $\mathbf{r}$ from the center of the atom be $\rho(\mathbf{r})$. Consider the wave scattered at position $\mathbf{r}$ in a direction $\mathbf{s}$, relative to the wave scattered at the center of the atom. The resulting total scattered wave depends on the phase difference between the scattered waves.

Define the direction of the incident wave by a vector $\mathbf{s_0}$ and the direction of the scattered wave by a vector $\mathbf{s}$. For simplicity in later equations let $|\mathbf{s_0}|=|\mathbf{s}|=1/\lambda$ where λ is the wavelength of the incident radiation.

From Fig. 2.3 follows that the path difference between ray 1 and ray 2 is p-q.

$$p = \lambda \mathbf{r} \bullet \mathbf{s_0} \quad (2.9) \qquad\qquad q = \lambda \mathbf{r} \bullet \mathbf{s} \quad (2.10)$$

Phase difference

$$\Delta \varphi = \frac{2\pi}{\lambda}(p-q) = 2\pi \mathbf{r} \bullet (\mathbf{s_0} - \mathbf{s}) = 2\pi \mathbf{r} \bullet \mathbf{S} \quad (2.11)$$

where $\mathbf{S} = \mathbf{s_0} - \mathbf{s}$.
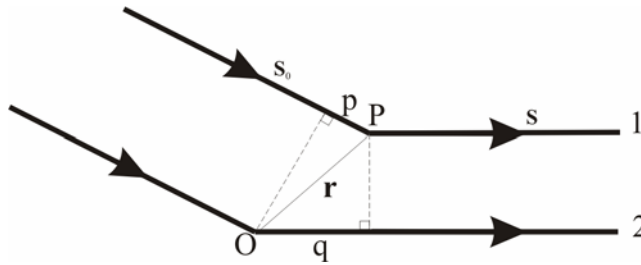
Figure 2.3: Scattering at a point P relative to an origin O.

Therefore, the total wave scattered by a small unit of volume $d\mathbf{r}$ at a position $\mathbf{r}$ relative to the wave scattered at the origin is:

$$F(\mathbf{S}) = V(\mathbf{r})e^{2\pi i \mathbf{Sr}} d\mathbf{r} \quad (2.12)$$

Hence the total wave scattered by the atom is calculated by summing the individual contributions over the volume of the atom

$$f(S) = \int\limits_{\substack{Vol.of \\ atom}} V(\mathbf{r})e^{2\pi i \mathbf{Sr}} d\mathbf{r} \quad (2.13)$$

This expression represents the *atomic scattering factor*.

Now we will derive the expression for the scattering by a molecule. Let us consider atom 1 in Fig. 2.4 which is at a distance $\mathbf{r_1}$ from the origin. This shift in origin from the center of the atom means that the distance $\mathbf{r}$ in the equation 2.13 for the scattering by an atom becomes $\mathbf{r}+\mathbf{r_1}$.

Hence the scattering by atom 1 relative to the new origin is

$$F(\mathbf{S})_1 = \int_{\substack{Vol.of \\ atom}} V(\mathbf{r})e^{2\pi i \mathbf{S}(\mathbf{r}_1+\mathbf{r})}d\mathbf{r} = f_1 e^{2\pi i \mathbf{S}\mathbf{r}_1} \qquad (2.14)$$

where

$$f_1 = \int_{\substack{Vol.of \\ atom}} V(\mathbf{r})e^{2\pi i \mathbf{S}\mathbf{r}}d\mathbf{r} \qquad (2.15)$$

Similar expressions may be deduced for atoms 2, 3 and all the other atoms in the molecule.

The total wave scattered by the molecule is given by the vector sum of the individual contributions from each of the atoms
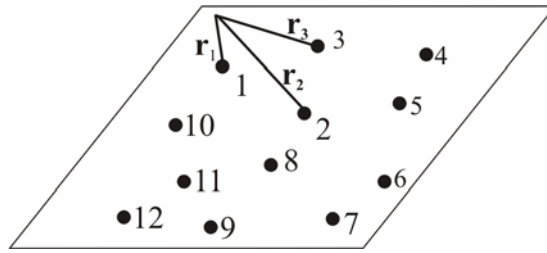
Figure 2.4: The positions of atoms in a molecule.

$$F(\mathbf{S})_{mol} = F(\mathbf{S})_1 + F(\mathbf{S})_2 + ... + F(\mathbf{S})_N = \sum_{j=1}^{N} f_j e^{2\pi i \mathbf{S}\mathbf{r}_j} \qquad (2.16)$$

Equation 2.16 represents the *molecular structure factor*, or the molecular transform.

Since a general goal of the scattering experiment is to calculate the electron (nuclei) density of the molecule, formula 2.16 may be rewritten:

$$F(\mathbf{S}) = \sum_{j=1}^{N} f_j e^{2\pi i \mathbf{S}\mathbf{r}_j} = \int_{\substack{Vol.of \\ molecule}} \rho(\mathbf{r})e^{2\pi i \mathbf{S}\mathbf{r}}d\mathbf{r} \qquad (2.17)$$

where $\rho(\mathbf{r})$ is the electron (nuclei) density.

By multiplying both sides by $e^{-2\pi i \mathbf{S}\mathbf{r}}$ and integrating over the volume of the molecule it may be shown that

$$\rho(r) = \int_{\substack{Vol.of \\ molecule}} F(\mathbf{S})e^{-2\pi i \mathbf{Sr}} d\mathbf{r} \qquad (2.18)$$

If a crystal is used, the formula 2.17 can be rewritten:

$$\left|F_{cryst}\right|^2 = \left|F_{mol}\right|^2 \left(\frac{\sin^2(\pi N_x \mathbf{Sa})}{\sin^2(\pi \mathbf{Sa})} \frac{\sin^2(\pi N_y \mathbf{Sb})}{\sin^2(\pi \mathbf{Sb})} \frac{\sin^2(\pi N_z \mathbf{Sc})}{\sin^2(\pi \mathbf{Sc})} \right) \qquad (2.19)$$

where $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ are lattice vectors; N_x, N_y, N_z are a number of unit cells in the x, y, z direction respectively.

Equation 2.19 described the intensity distribution along the reciprocal-lattice lines in Fourier space. The crystal structure factor has distinct maxima when Laue conditions are obeyed.

In the case of 2D crystals (N_x, $N_y \to \infty$, N_z=1), the intensity distributions are δ-functions in the $\mathbf{a}^*$, $\mathbf{b}^*$ plane with a constant distribution in the third direction.

For multilayered crystals the number of layers in the z direction is limited. It results in the elongated density profiles in reciprocal space (Fig. 4.10).

If the periodicity in a certain direction is increased, the lattice transform in this direction is "shortening". Therefore, for a real 3D crystal (N_x, N_y, $N_z \to \infty$) the lattice transform is a lattice of δ-functions which have non-zero values at the reciprocal lattice points.

2.3 Interaction of X-rays and Electrons with Matter

The theory described in subchapters 2.1 and 2.2 is valid for any type of wave scattered elastically. In fact, the choice of the appropriate wavelength and nature of the incident beam depends on the goal of the actual experiment. Electron and X-ray beams are widely used for biological structural studies, since they may provide atomic and quasi-atomic resolution. When an electron or photon interacts with matter two different types of scattering may occur:

1) In an *elastic scattering* event the energy is (by definition) conserved and k_0=k;

2) Inelastic scattering is a process by which the particle loses energy and, therefore, $k_0 \neq k$.

This paragraph summarizes different types of interactions of X-rays and electrons with matter.

2.3.1 Photon-Specimen Interactions

There are four possible processes that may occur when photons interact with matter (see also Fig. 2.5):

1) The **photoelectric effect** (Fig. 2.5a) is characterized by a photon giving all of its energy to a bound electron, ionizing the atom in a single interaction. The photoelectric effect (see also Fig. 2.7) is therefore observed only for photon energies higher than the binding energy of the electrons. The photoelectric effect is the dominant interaction of photons with matter for those photons whose energy falls approximately in the range 1 keV< E_γ < 0.5 MeV (1.2 nm >λ> 2.5 pm). The cross section for this process is proportional to $\dfrac{Z^4}{E_\gamma^3}$.

2) **Compton scattering** (Fig. 2.5b) is inelastic scattering of a photon by an electron. In the simplest case, a free electron at rest is hit by a photon that scatters off at some angle to its original path. By the conservation of momentum and energy, the electron must recoil in a specific direction with a specific energy. This energy must be provided by a reduction in the energy of the photon (increase in its wavelength). The probability of Compton scattering is proportional to $\dfrac{1}{E_\gamma}$, at energies around 1 MeV (λ=1.2 pm) it is the most common interaction.
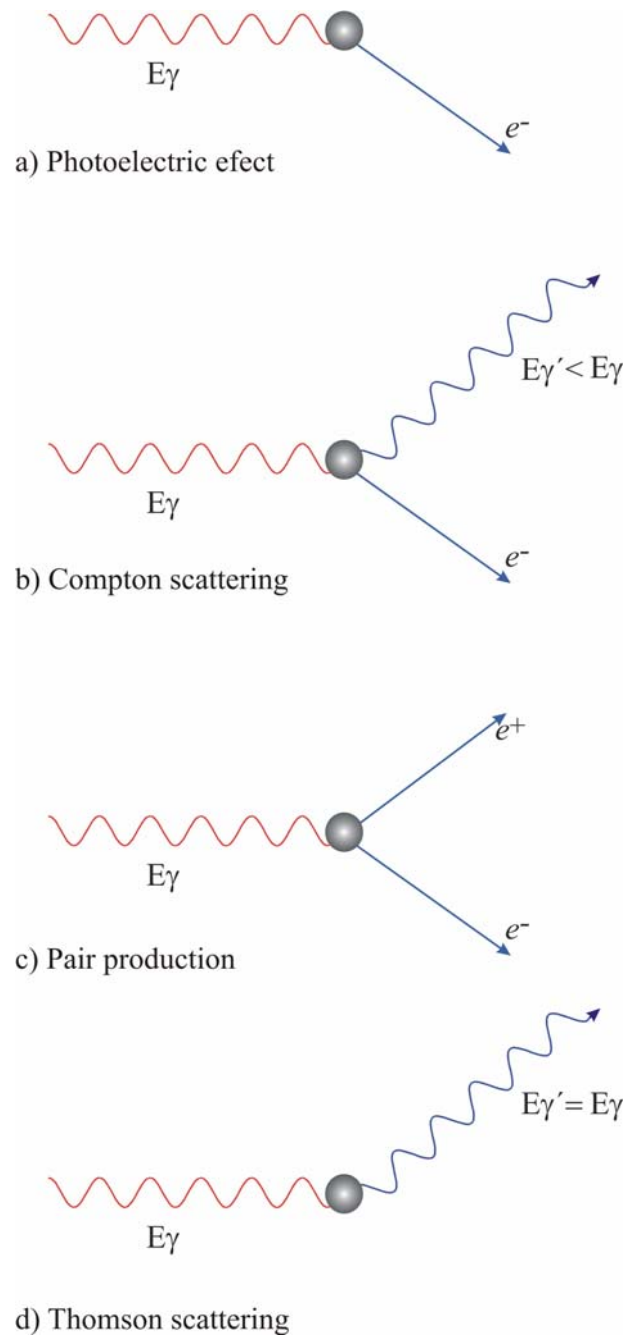
Figure 2.5: Photon-specimen interactions : a) Photoelectric effect; b) Compton scattering; c) Pair production; d) Thomson (elastic) scattering.

3) Pair production (Fig. 2.5c). At photon energies above about 1 MeV another process becomes energetically available when matter is struck by the photon: the creation of particle-anti-particle pairs. The lightest particle is the electron. To make

an electron and its anti-particle, a positron, twice the rest mass energy of an electron is required (since the positron as anti-particle has the same mass as an electron). The pair production effect occurs when $E_\gamma > 2m_e$ and is the dominant interaction for photons of energy above 10 MeV (λ=0.12 pm). The probability of pair production is proportional to $E_\gamma Z$.

4) **Thomson scattering** (Fig. 2.5d). Thomson scattering, also known as coherent, or classical scattering, is an elastic scattering effect. It occurs when the X-ray photon in a energy range 0.01 keV$<E_\gamma<5*10^2$ keV (12 μm $>\lambda>$ 2.5 pm) interacts with an electron. The basis of the physical effect is the interaction of a wave with a particle, which is purely described by classical mechanics. The assumption made in the Thomson analysis is that the electron is free and responds passively to the electromagnetic wave in which it is placed. When an electromagnetic wave is incident on a charged particle, the electric and magnetic components of the wave exert a Lorentz force on the particle, setting it into motion. Since the wave is periodic in time, the motion of the particle is also periodic. According to the classical electromagnetic theory, any accelerated charge acts as a source of radiation. It means, that the electron emits waves, known as scattered waves. More exactly, energy is absorbed from the incident wave by the particle and reemitted as electromagnetic radiation. The fact that the acceleration of the electron is proportional to the electric field vector of the incident wave implies that the phase relationship between the incident electric field and the scattered wave is well defined. Any system in which phase relationships are well defined is described as a coherent system, and so Thomson scattering is often referred to as coherent scattering. This implies that the overall effect of an assembly of electrons may be calculated by using the "Principle of Superposition" in which the total amplitude of the scattered wave is the sum of individual amplitudes scattered by each electron. Therefore diffraction effects may be observed and so Thomson scattering is a valid model for the diffraction of X-rays by crystals.

2.3.2 Electron-Specimen Interactions

Electrons are charged particles and the interaction of electrons with the specimen is dominated by the Coulomb interaction. As in the case of X-rays one can distinguish elastic and inelastic processes.

In **elastic scattering**, electrons are deflected by the electric field of the atomic nucleus of the specimen atoms. In this process the kinetic energies and momenta of the colliding particles remain constant. High resolution structural information can be obtained from elastically scattered electrons because the Coulomb potential of the positively charged atomic nucleus is well localized. Elastic scattering is most important for image contrast and electron diffraction.

Inelastic scattering results from the interaction between the incident electrons and the bound electrons of the sample (see Fig. 2.6). These interactions result in an energy loss of the incident electrons by the excitation of different energy states of the specimen. Inelastic scattering processes are not localized as well as the Coulomb potential of the nucleus, leading to smaller scattering angles and are generally not used for high resolution structural studies. Instead, inelastic scattering provides analytical information about the chemical composition and state of the sample. Inelastic interactions consists of six major inelastic scattering effects:

1) **Phonon excitation** (heating). Much of the energy deposited into the specimen by the incident electron beam is transferred to the solid body by the excitation of lattice oscillations (phonons), that is, heat. If the specimen is a good thermal conductor, the specimen and the specimen holder will serve as an effective heat sink, and significant heating will not occur. For nonconductors, heating can be more significant. In some cases, beam-induced heating is sufficient to cause phase transformations or recrystallization in specimens**.**

2) **Plasmon excitation.** The second type of inelastic scattering is the interaction with valence and conduction band electrons. These electrons respond with a collective oscillation (correlated motion) to the excitation by the primary electron. The quantized pseudo-particle describing this collective oscillation is called the

plasmon. Its energy (lost by the primary electrons) is in the 5-30 eV range. It is a frequent scattering event resulting in a strong peak in the EELS[5] spectrum. The same process is responsible for the very short mean free path of the beam electrons and the multiple scattering that smears out the characteristic EELS edges when the sample thickness increases.
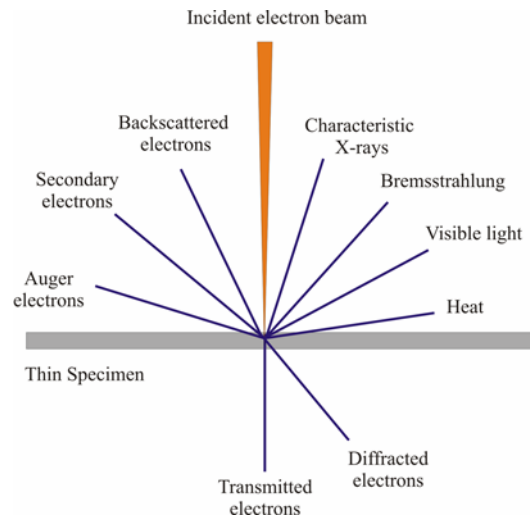
Figure 2.6: Effects occurring upon electron bombardment of a material.

3) **Secondary electrons.** Interaction of the primary electrons with outer-shell electrons leads to small energy losses (typically a few eV) and scattering into small angles. The excited electron may be emitted as a secondary electron. Since the amount of energy given to these secondary electrons as a result of the interactions is small, they have a very limited free path length in the sample (a few nm). Because of this, only those secondary electrons that are produced within a very short distance of the surface are able to escape from the sample.

4) **Characteristic X-ray and Auger electrons.** Another way in which a beam electron can interact with an atom is by the ionisation of an inner shell electron. The resultant vacancy is filled by an outer electron, which can release its energy either via an Auger electron or by emitting a photon. This produces characteristic lines in the X-ray spectrum corresponding to the electronic transitions involved.

---

[5] Electron energy-loss spectrometry (EELS) is the analysis of the energy distribution of electrons that have interacted inelastically with the specimen.

Since these lines are specific for a given element, the composition of the material can be deduced. This can be used to provide quantitative information about the elements present at a given point of the sample, or alternatively it is possible to map the abundance of a particular element as a function of position.

5) **Cathodoluminescence.** When the beam is incident on a semiconducting sample, the high energy primary electrons can promote valance band electrons into conduction band via the process of impact ionisation. The resultant electron-hole pairs can then recombine, either by radiative or non-radiative mechanisms. Spectroscopy of the visible and near-visible light emitted from the sample allows to obtain information about the radiative recombination processes.

6) **Bremsstrahlung** (continuum X-ray generation) is the result of a continuous energy-loss process generated by an electron penetrating into the specimen and undergoing collisions with the atoms in it, resulting in electromagnetic radiation.

2.4 Electrons versus X-rays

In the following we will discuss advantages and disadvantages of using electrons versus X-rays in diffraction experiments in structural biology.

In general, in diffraction experiments information from elastically scattered electrons (X-rays) is used, while inelastic scattering results in one component of the background noise. During the inelastic event an electron (photon) loses energy and this energy is deposited into the sample. This deposition process causes radiation damage and results ultimately in the destruction of the sample. All biological materials are very sensitive to radiation damage and, therefore, only a limited amount of energy can be absorbed by the specimen without alteration of the internal structure.

To record an interpretable signal, a given number $N_{el,signal}$ of elastic events must occur during the data collection time t. However, inevitably during the same time t N_{inel} inelastically events occur as well. The total maximal energy which can be absorbed by the specimen is:

$$E_{total} = \left(\frac{\sigma_{inel}}{\sigma_{el}} \right) E_1 N_{el,max} \qquad (2.20)$$

where σ_{inel} and σ_{el} are inelastic and elastic cross-sections; E_1 is energy deposited per inelastic event and $N_{el,max}$ is the maximal number of elastic events before the sample is destroyed. If $N_{el,max} < N_{el,signal}$, no information can be extracted from the recorded data. As outlined before, σ_{inel}, σ_{el}, E_1 depend on the energy of electrons or X-rays.

It can be seen from Table 2.1 (adopted from Henderson (27)) that, for example, 30 Å X-rays deposit 20 times more energy per inelastic event than an electron in the energy range of 80-500 keV. Furthermore, the total energy deposition into the specimen per elastic event is 10^4 times higher for 30 Å X-rays than for electrons.

Table 2.1: Energy deposited in biological specimen per useful scattering event
(adopted from Henderson (27))

		Electrons	X-rays	
		80-500 keV	1.5 Å	30 Å
Ratio (inelastic/ elastic) scattering events		3	10	10^3-10^4
Energy deposited per inelastic event		20 eV	8 keV	400 eV
Energy deposited per elastic event[6]		60 eV	80 keV	400 keV
Energy deposited relative to electrons	inelastic	1	400	20
	elastic	1	10^3	10^4

---

[6] $E_{el} = E_{in}*r$, where E_{el} is energy deposited per elastic event; E_{in} is energy deposited per inelastic event and r is a ratio (inelastic/elastic) scattering events.

Figure 2.7 shows a comparison between electrons, X-rays and neutron scattering cross-sections for typical organic materials. The inelastic and elastic cross-section for both electrons and X-rays vary steeply with energy (adopted from Henderson (*27*)).

In the case of X-rays, the inelastic cross-section excluding edge effects varies with λ^3 whereas the total elastic cross-section tends towards a λ^2 variation (*27*) at the high energy end of the spectrum (Fig 2.7). Therefore, at longer wavelength, relatively more X-ray quanta are absorbed than scattered. However, these quanta deposit less energy in the
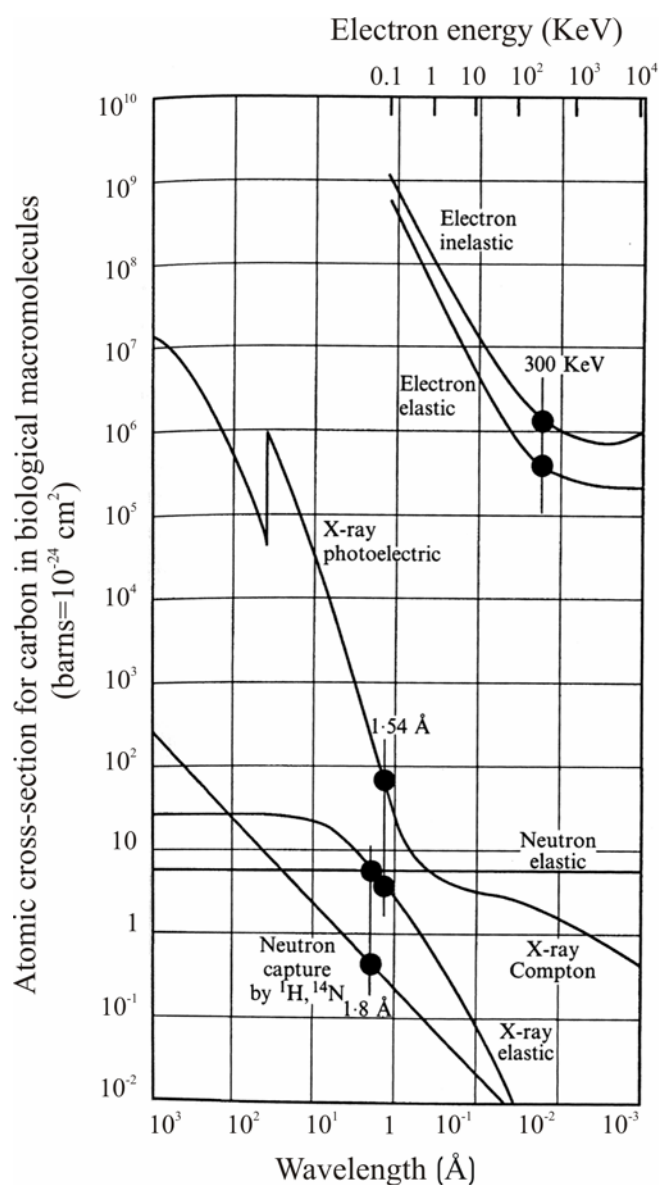
Figure 2.7: Comparison of the atomic cross-sections for electrons, X-rays and neutrons as a function of wavelength of the particle or quantum. "•" denotes typical energies used for data collection (Adopted from (*27*))

specimen. The elastic cross-section becomes constant at longer wavelengths and decreases at shorter wavelengths.

In the case of electrons, the elastic and inelastic cross-sections decrease at higher energy (shorter wavelength) until the nuclear displacement process becomes important, and then the inelastic cross-section increases again (at about 1.5 MeV, $\lambda=8.3*10^{-13}$ m). The cross-sections for elastic and inelastic vary in a roughly parallel manner with energy, being inversely proportional to the electron velocity. Below 100 keV ($\lambda= 1.2*10^{-11}$ m), both cross-sections vary strongly with energy tending towards a λ^2 dependence at low energy. However, above 200 keV ($\lambda=6.2*10^{-12}$ m), as the electron speed approaches c, the cross-sections are flattening out.

2.5 Electron Crystallography

R. Henderson (*27*) has analysed the potential and limitations of electrons and X-rays for atomic resolution structure determination of unstained[7] biological molecules. He discussed that radiation damage is a severe problem in structural studies because energy is deposited into the specimen when an inelastic event occurs. It means that a single organic molecule can be destroyed by a dose of X-rays or electrons far less than that required to determine its structure at high resolution. Therefore, to reduce beam damage, the total X-ray (electron) dose must be minimized. Consequently, the signal-to-noise ratio is very low and averaging over many molecules must be performed to determine molecular structures at high resolution.

If electrons are used, there are two ways to perform particle averaging. In the first method, the so-called the single particle reconstruction, three dimensional structural information can be obtained by combining different projections of single proteins or protein complexes.

---

[7] As it is known, water is the most abundant component of biological material. Therefore, a biological sample must be protected due to the vacuum conditions inside the column of the electron microscope to preserve as best as possible the native structure of the specimen without water. On of the most successful methods is *negative staining*, in which the specimen is embedded in a thin layer of a heavy atom salt. In another method, so-called cryo-electron microscopy, the sample remains *unstained* and is frozen in a very thin layer of ice to avoid damage of the sample during drying.

Unfortunately, the effectiveness of this approach is limited for molecules smaller than about 10^5 Da (*27*)[8] . Furthermore, maps at resolution of only 10-20 Å can be obtained. As a result of these considerations, the second averaging method which uses crystals and provides high resolution data, is often the method of choice.

In crystallography, a crystalline sample is used to perform particle averaging. Nowadays, X-ray crystallography is a routine method which utilizes 3D crystals. However, due to a relatively small elastic cross-section for X-rays, typical sizes of protein crystals are about 100-200 μm in each direction for X-ray beams generated by a rotating anode and about 15-30 μm for a microfocused synchrotron beam. Micrometer-size protein crystals, microcrystals, cannot be used for X-ray analysis, yet.

If the number of scatterers (i.e. protein molecules) is decreased, due to the smaller crystal size, the probability of a particle being scattered must be increased in order to obtain a measurable signal. This could be achieved by using electrons, without increase of the scattering cross-section by a factor of $5*10^4$ as compared to X-rays (see Fig. 2.7) (*27*). This results from the different physical processes of interaction with the scattering center by electrons and X-rays. The mean-free-path is reduced to about 100 nm allowing (and requiring) samples to be small. In principle, single protein complexes can be visualized, so that "crystals" could now be as small as one single molecule (*28*).

The possibility to work with such small crystal size is a big advantage for electron crystallography, but the mathematical interpretation of the diffraction data is more complicated. Firstly, with the high scattering cross-section, the sample may be too thick for interpretation by the kinetic scattering theory (for details see (*21*)). Increased thickness of the sample leads to dynamic scattering effects, which affect diffraction intensities and causes a break in Friedel symmetry. This would lead to uninterpretable data for a 3D reconstruction. As shown by Glaeser and Ceska (*29*), the dynamic scattering effect is present even in thin biological samples such as monolayered bacteriorhodopsin crystals.

---

[8] In principles, structures can be determined for any molecules of size larger than about 10^5. However, in practice molecules larger than 10^6 Da are already within accessible resolution range (*27*).

The effect of dynamic scattering on diffraction data from multilayered protein crystals was analysed by Vossen (*21*). He came to the conclusion that dynamic scattering does not influence the 3D reconstruction at a resolution of about 4-6 Å in a densely packed sample up to 5 layers (30 nm of the typical membrane protein bacteriorhodopsin which was used as the model) even when low energy electrons (120 keV) were used for data simulation. As the electron energy increases, dynamic scattering decreases. If higher electron energy was used (200-300 keV), interpretation of the reconstructed density at a resolution level on the order of 4-8 Å was possible for typical, less densely packed multilayered crystals of up to 15 layers. It was shown that symmetrization of the diffraction amplitudes serves as a first order correction mechanism of the dynamic scattering effects.

These theoretical simulations indicate the electron scattering data from a crystal thinner than about 50-100 nm can be adapted for use with X-ray data processing algorithms in order to obtain high resolution structural information.

Chapter 3

Protein Crystallization: Art or Science?

As we discussed in the previous chapter, X-ray and electron diffraction can be used to determine a protein structure. Unfortunately, a biological specimen is extremely sensitive against radiation damage. It means that a single protein molecule is destroyed by a dose of X-rays or electrons in less time than that required to determine its structure (*27*).

In the near future, free-electron lasers (FEL) will provide laser-like hard X-ray beams with ultrashort X-pulses. These pulses might be intense and short enough for diffraction to occur before the molecules are destroyed by a Coulomb explosion. Thus, FELs may allow to observe an image of a single big protein molecule with a single X-ray pulse. However, at the present time, a signal from many molecules is required for structure determination. The limitation of the strong beam damage is overcome by irradiation at a low dose and averaging over many molecules for detection.

In *Electron Microscopy* there are two alternatives to perform signal averaging: the use of single particles, whereby images of single particles are grouped on the basis of similar features and thus, presumably, similar orientations (*30*); or the use of crystals, which must be very thin to allow transmission of the electron beam and facilitate interpretation of the data. The single particle approach has the advantage that crystals are not required, but the effectiveness of correlation analysis for determining orientations is problematic for smaller molecules (less than about 10^6 Da). As a result of these considerations, electron crystallography is often the method of choice for complexes smaller than about 10^6 Da.

Among all current EM approaches only Electron Crystallography (EC) has the potential to provide atomic coordinates. Nowadays conventional EC utilizes only 2D crystals. In contrast to X-ray analysis, which deals with any protein, EC has mainly been used with membrane proteins because they often occur naturally in the form of 2D crystals embedded in lipid bilayers. This offers an opportunity to visualize the structure of the proteins easily. Over the past few years dozens of proteins have been crystallized in 2D and studied at various levels of resolution. Although the majority of protein structures were obtained at sub-atomic resolution (6-8 Å, for review see (*31*)), only five were solved at near atomic resolution which is required for understanding the molecular mechanism of the protein's function (*32-36*).

In the case of **X-rays**, at the present time, analysis of single particle images cannot be applied due to absence of a high resolution X-ray microscope or X-FEL. Therefore, only diffraction experiments can be performed. To do signal averaging, diffraction patterns must come from molecules that are identically orientated. The sample, which has a perfectly ordered internal structure, is a crystal. As we discussed in Chapter 2, in contrast to electrons X-rays have a much smaller scattering cross-section. It means that the sample must be large enough to provide a high probability of the scattering event. Consequently, a large, well-ordered 3D protein crystal must be used in the X-ray diffraction experiment.

It is obvious that the key for a successful X-ray analysis is to grow suitable crystals. However, the generation of crystals of sufficient size (typically at least 100 μm in all dimensions) is usually a time-consuming and labour-intensive process that relies only on empirical screening studies. At the present time, there is no theory to predict the conditions for crystal growth. Although X-ray crystallography is currently the predominant source of structural analysis, it is usually limited to water soluble proteins. Only a small number of membrane protein structures have been determined by X-ray analysis (*37-41*). Membrane proteins often form aggregates rather than well-ordered 3D crystals. To prevent proteins from aggregations and to keep them soluble in aqueous solutions, detergents must be used. However, using detergents extends the already tedious screening procedure significantly.

3.1 General Principles of Crystallization Theory

Crystallization can be viewed as a special case of phase separation on a thermodynamically non-ideal mixture. The material has to be separated from another phase (organic solution, melt, amorphous metal, aqueous protein solution) and should form a distinct and well-ordered crystalline solid phase. The crystallisation process of molecules of any compound from its solution represents a reversible equilibrium phenomenon, driven by the minimisation of the free energy of the system (*42*). A solution, in which the molecules are fully solvated, corresponds to the system at equilibrium where its free energy is minimised. If more molecules are added to the solution, or water molecules removed, the system undergoes internal changes until the point is reached where there is insufficient solvent to maintain full hydration of the molecules. Under these conditions (the so-called "supersaturated state") the system is no longer at equilibrium. Therefore, it will be thermodynamically driven toward a new equilibrium state with a corresponding new free energy minimum. Individual molecules lose rotational and translational freedom by forming many new stable non-covalent chemical bonds, thus minimising the free energy of the new system. Crystallisation is known to lower the free energy of proteins by approximately 3 - 6 kcal/mole relative to the dissolved state in solution (*43*). This aggregation results in partitioning of the molecules between soluble and solid phases. The solid phase can appear in a state of either amorphous precipitate or crystal nuclei. Amorphous precipitates are usually more favourable kinetically, and so they tend to dominate the solid phase and inhibit crystal formation.

The principles of crystal growth have been the subject of intense investigation for a number of years. As a consequence, the theoretical and practical aspects of crystallisation of molecules like salts or small organic compounds are nowadays well established. Over the past decade these approaches have been applied for crystallisation of macromolecules, like proteins, DNA and RNA. Protein crystallization is now seen as very similar to that of most other chemical systems. The only notable differences for protein crystallization systems are the requirement of very high supersaturation (relative supersaturation about 100-500%) to induce nucleation and subsequent crystal growth and the high solvent content (~50%) of the crystals.

The basic strategy of all crystallization processes is to bring the system into a state of limited degree of supersaturation by modifying the properties of the solvent by changing the precipitating agent concentration, or by altering some physical properties such as temperature, etc. Three stages of crystallisation are common to all systems: nucleation, growth, and cessation of growth.

The nature of the crystal nuclei is largely a mystery. It is not known whether they are initially ordered, ordered after a certain degree of internal re-arrangement, or whether they form by the coalescence of random subnuclear clusters or by strict monomer or oligomer addition. It is believed that in the crystal nuclei, molecules show the tendency to make interactions with one another in all three directions in a periodic manner (*44*). A different scenario occurs when amorphous precipitates form, as they are characterised by random intermolecular interactions. These unstructured aggregates tend to form over a much faster time scale than regular crystals.

The formation of crystal nuclei from supersaturated solutions does not necessarily imply the subsequent formation of macroscopic crystals. In fact, the nucleus must first exceed a specific size, called the critical size, defined by the ratio between the surface area of the nucleus and its volume (*45*). Once the critical size is exceeded, the nucleus is capable of further growth, given the availability of protein molecules in solution and suitable physico-chemical experimental conditions. If the nuclei are smaller than the critical size, spontaneous dissolution will occur.

A classical explanation of crystal nuclei formation and growth is given by the two-dimensional solubility diagram shown in figure 3.1. The solubility curve divides the phase space into two areas - the undersaturated and supersaturated zones. Each point on this curve corresponds to a concentration at which the solution is in equilibrium with the crystallizing agent. These correspond to the situation either at the end of the crystal growth process from a supersaturated solution or to a situation when crystal dissolution occurs in an undersaturated solution. In the area under the solubility curve, the solution is undersaturated and the crystallisation will never take place. Above the solubility curve lies the supersaturation zone. Here, for a given concentration of precipitating agent, the protein concentration is higher than that at equilibrium. Depending on the kinetics to reach
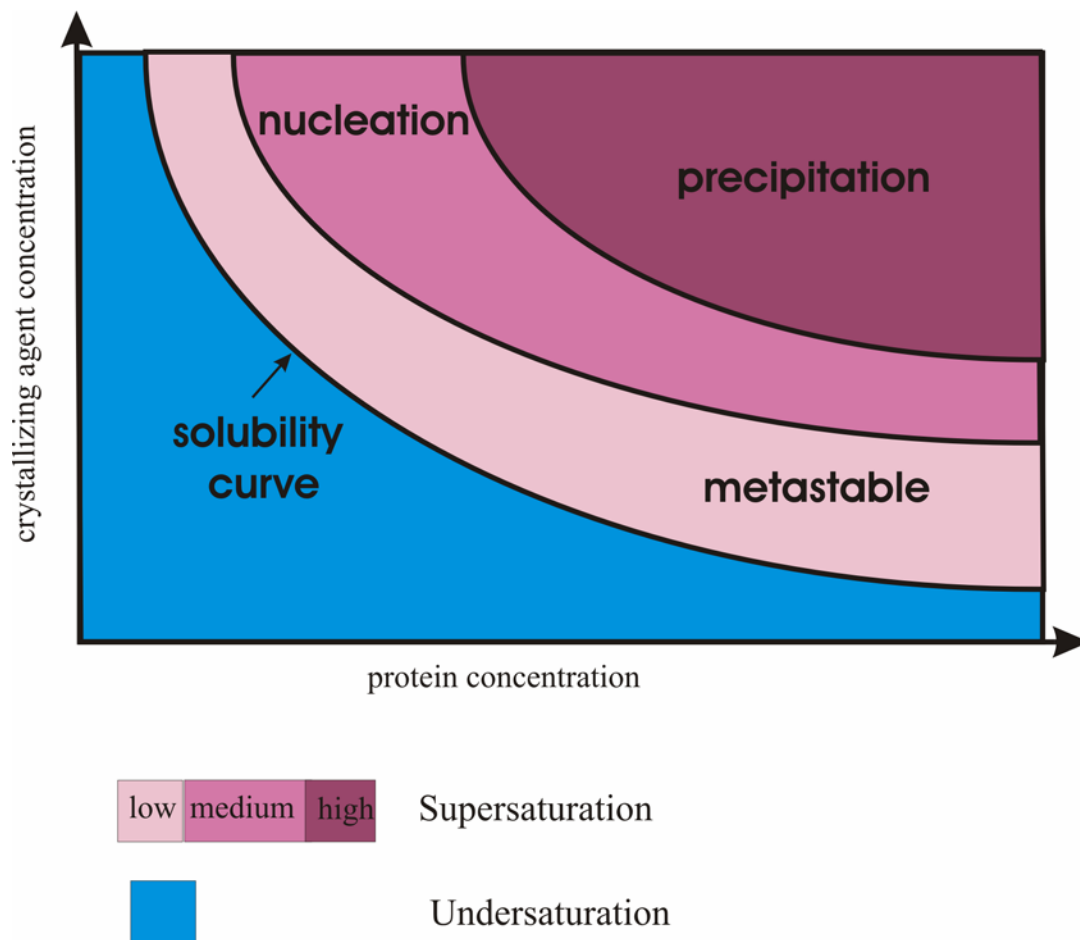
Figure 3.1: Crystallization phase diagram. Schematic representation of a two dimensional phase diagram, illustrating the change of protein concentration against crystallizing agent concentration. The phase space is divided by the solubility curve into areas corresponding to undersaturated (blue) and supersaturated (red) state of a protein solution. The supersaturated area compromises of the metastable, nucleation and precipitation zones.

equilibrium and the level of supersaturation, this region may itself be subdivided into three zones:

1. The precipitation zone, where the excess of protein molecules immediately separates from the solution to form amorphous aggregates.

2. The nucleation zone, where the excess of protein molecules aggregates in a crystalline form. Near the precipitation zone, crystallisation may occur as a shower of microcrystals, which can be confused with precipitate.

3. In the metastable zone a supersaturated solution may not nucleate for a long period, unless the solution is mechanically shocked or a seed crystal introduced. To grow well-ordered crystals of large size, the optimal conditions would have to begin with the formation of a preferably single nucleus in the nucleation zone just beyond the metastable zone. As the crystals grow, the solution would return to the metastable region and no more nuclei would be formed. The remaining nuclei would grow, at a decreasing rate that would help to avoid defect formation, until equilibrium is reached.

Cessation of crystal growth can occur for many reasons. The most obvious is the decrease in concentration of the crystallising solute to the point where the solid and solution phases reach exchange equilibrium. In this case, the addition of more solute can result in continued crystal growth. However, some crystals reach a certain size beyond which growth does not proceed, regardless of solute concentration. This may be the result of either cumulative lattice strain effects or poisoning of the growth surface.

Poisoning of growing crystals occurs when foreign or damaged molecules are incorporated into the growing crystal face, resulting in successive defects, which interrupt the crystal lattice. An example of this might be the incorporation of a proteolytically damaged protein onto the face of an otherwise perfect protein crystal. If the damaged molecule is unable to form the same lattice contacts with newly added molecules, as would the perfect protein, then its incorporation will cause local defects in the growing crystal. As crystals grow larger, the likelihood of incorporation of defective molecules into the lattice increases. Sato and co-workers have used laser scattering tomography to visualise lattice defects in large orthogonal crystals of hen egg white lysozyme (*46*). Their results demonstrated the occurrence of defects not only at the surface, but also within the bulk of the crystal itself.

3.2 Variables Affecting Protein Solubility

Although biological macromolecule systems share the same fundamental mechanism and pathways of crystal growth as small molecules, the magnitude of the underlying kinetics and thermodynamic parameters regulating the process differ dramatically. The major contributing factors include:

- the complexity and high degree of mobility at the surface and as a whole for large molecules;
- the electrostatic nature of the macromolecule surface;
- the relatively low chemical and physical stability of macromolecules (unfolding, hydration requirements, temperature and mechanical sensitivity);
- molecule specific factors (such as prosthetic groups and ligands). The above aspects are also reflected in an enhanced protein crystal disorder, which is one of many factors limiting the resolution attainable in a diffraction experiment.

The other crucial difference between protein and small molecule crystallisation is that protein crystals usually require extremely high levels of supersaturation for nucleation. Although high levels of supersaturation are necessary to promote nucleation of protein crystals, these conditions are normally far from ideal for crystal growth. In fact, when supersaturation is achieved, the competition between crystal nuclei and amorphous precipitation becomes severe. Additionally, biomacromolecular crystals have to be grown from chemically rather complex aqueous solutions, that in turn increase the number of parameters that need to be controlled for successful protein crystallisation. Consequently, the optimal conditions of protein crystal growth are virtually impossible to predict *a priori*.

As in any crystallization, the production of protein crystals requires bringing the protein into a supersaturated liquid state. The degree of supersaturation determines the rate of nucleation as well as the crystal growth rate. The degree of supersaturation is defined by the equilibrium solubility of the protein. The principal question is: Which factors can be changed – in a reasonably controlled fashion – to get a protein solution into a metastable zone. Different variable parameters are used for achieving supersaturation of

macromolecules in order to induce crystallization. The thermodynamic variables such as pH, composition and temperature and are typically used to alter the solubility of proteins.

3.2.1 Effect of pH on Protein Solubility

A protein is the charged molecule and its net charge can be modified by adding (or withdrawing) protons into the solution. Generally, the protein is more soluble the more net charge it contains. The solubility minima correspond well with the isoelectric point (pI, the pH at which the net charge of the protein is zero). Unfortunately, the correlation of pI and the actual pH of crystallization is weak. Therefore, different ranges of pH have to be tested to find optimal crystallization conditions.

The majority of the proteins tend to crystallise at physiological pH (around pH 6.5). However, some proteins have been crystallised at extreme values of pH. For some proteins, crystallisation occurs over a very broad range of pH with little variation in crystal morphology. However, it is far more typical for crystallisation to occur over a narrow range. In this case, protein solubility is changed dramatically as pH is altered by 0.5 pH units.

3.2.2 Effect of Electrolyte on Protein Solubility

The ''salting out'' effect has been used for protein purification for a long time. The basic idea is that ''salting'' the solution (adding electrolyte) causes the protein to come ''out'' as a separate solid phase. The effect of the electrolyte on protein solubility is usually described as salting out or salting in. Salting in implies an increase in the protein solubility as the electrolyte is added to the solution. This "salting in" effect can be explained by considering the protein as an ionic compound. According to the Debye-Hückel theory for ionic solutions, an increase in the ionic strength lowers the activity of the ions in the solution and increases the solubility of ionic compounds. Alternatively, one can regard the

"salting in" effect as the result of competition between charged groups on the surface of the protein molecule and the ions in the solution. In the absence of ions, the protein precipitates by Coulombic attraction between opposite charges on different protein molecules. If ions are added, they shield the charged groups on the protein and increase its solubility. A "salting out" effect occurs when the ionic strength is increased to the point where salt ions and protein molecules compete for solvent molecules i.e. water, to maintain their hydration layers and therefore their solubility. To fulfil their electrostatic requirements, the protein molecules begin to self-associate.

3.2.3 Effect of Antisolvents on Protein Solubility

Other common chemical additives to reduce protein solubility are small, polar organic molecules such as methanol, ethanol, and acetone. Such water-miscible solvents are often called antisolvents. Antisolvent addition has multiple effects on the protein. Firstly, organic solvent binds water molecules and diverts them from their interaction with protein atoms. Secondly, it also lowers the dielectric constant of the medium, thereby enhancing Coulomb interactions between protein molecules and reducing their solubility. These effects decrease the capacity of the system to fully solvate the protein molecules and reduce the electrostatic shielding, which in turn forces the protein molecules out of the solution. Some of these agents should be used at low concentrations since they often destabilise protein molecules at high concentrations.

3.2.4 Effect of Soluble Synthetic Polymers on Protein Solubility

The use of high molecular weight linear polymers as precipitating agents was pioneered by Polson and co-workers, who tested a variety of polymers including polyethylene glycol, dextran, polyvinyl alcohol, and polyvinyl pyrrolidone (47). Polyethylene glycol (PEG) was found to be the most effective both in terms of precipitating ability and cost effectiveness (48). PEGs are produced in a variety of molecular weights, ranging from 200 to 1000000

Daltons. Like salts, PEGs compete with protein solutes for water. However, unlike salts, PEGs decrease the dielectric constant of the solution, which increases the effective distance over which protein electrostatic effects occur. This makes them particularly well suited for macromolecular crystallisation. PEGs of molecular weight 4000 or 6000 seem to be very effective for most protein systems. In general, the higher the molecular weight, the more effective the PEG is at reducing the protein solubility.

3.2.5 Effect of Temperature on Protein Solubility

Temperature is another important parameter in the crystallisation of a protein. Proteins have been documented to crystallize at a wide variety of temperatures. The vast majority of protein molecules are crystallised either at 4ºC or 18 - 22ºC (room temperature). Most proteins are denatured at temperatures above 40ºC, so temperature can be varied over only a limited region (i.e. 0–40ºC). The solubility of proteins in salt solutions tends to increase at low temperatures, whereas in PEG solutions protein solubility generally decreases with decreasing temperature. By increasing or decreasing precipitant or protein concentration, crystallisation should, at least in theory, be possible at either room temperature or 4ºC, although the kinetics of crystallisation can be expected to vary in accord with temperature.

3.3 Protein Crystallization Techniques for 3D X-ray Crystallography

A number of techniques have been developed for protein crystallization. Among them, the following three methods are frequently used: microbatch, vapour-diffusion and dialysis (figure 3.2) (*49*). Although supersaturation of a protein solution may be achieved by means of each of these techniques, the underlying principles of these methods vary.

Figure 3.2: Protein crystallization techniques. Schematic representation of a) microbatch; b) vapour-diffusion and c) dialysis crystallization. Adopted from (*49).
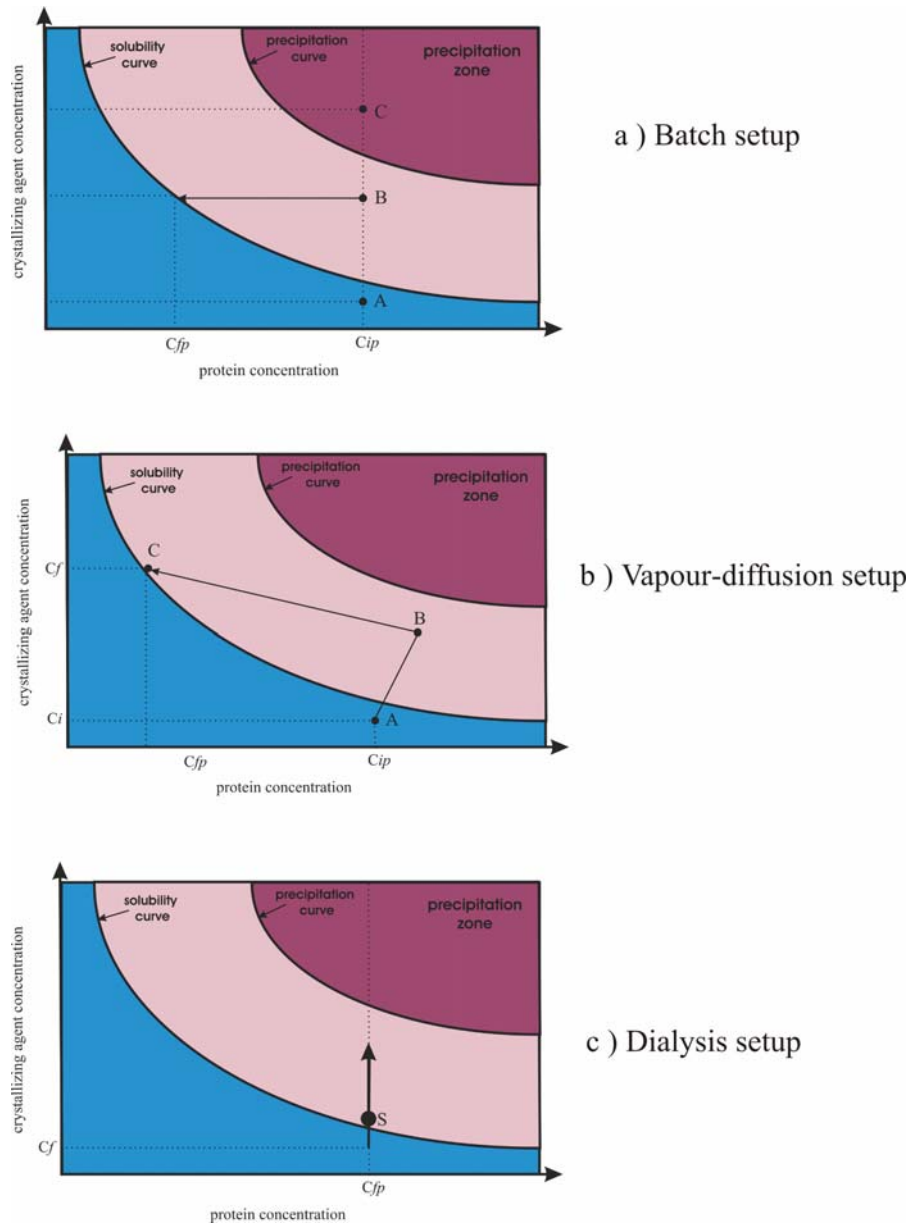
Figure 3.3: Various crystallisation setups explained in terms of phase diagrams. Schematic representation of solubility phase diagram and correlation between protein and crystallising agent concentrations in **a)** batch; **b)** vapour-diffusion and **c)** dialysis crystallisation experiments. C_{ip} and C_i are the initial concentrations of protein and crystallising agent respectively, C_{fp} and C_f are their final concentrations. Adopted from *(49)*.

The **microbatch method** is a variation of the simple batch crystallisation technique, where the concentrated protein solution is mixed with concentrated precipitant in a closed vessel to produce a final supersaturated concentration, which may eventually lead to crystallisation. This can be done with large amounts of solutions, and typically results in

larger crystals owing to the larger volumes of solute present. In the microbatch technique, smaller volumes (0.5-2 μm) can be used. The protein sample and precipitant solution are dispensed into the well of a plate and the well is covered with paraffin oil to prevent evaporation (figure 3.3a). During the incubation period, the concentration of a precipitant agent remains constant since evaporation is limited and, therefore, the volume of the drop remains the same during the experiment. On the other hand, the concentration of the protein changes on formation of either crystals or amorphous precipitant. If the concentration of precipitant agent is chosen in such a way that the solution is in an undersaturated state, crystallisation will never occur. In terms of phase diagram (figure 3.3a) this condition is indicated by point A. The protein will immediately precipitate if the starting point is located in the precipitation zone of the solubility diagram, point C. Therefore, only those points in the phase diagram (point B in figure 3.3a) which lie between the solubility and precipitation curves represent starting conditions for a successful crystallisation experiment. The main disadvantage of this method is that the equilibration occurs very rapidly, thus affecting the rate of crystal growth and consequently the quality of the obtained crystals. The second disadvantage is that the manipulation of the crystals from the drop covered by oil is very difficult. However, it allows the use of very small volumes of protein solution.

The **vapour-diffusion technique** utilises evaporation and diffusion of water between solutions of different concentrations. Typically, the protein solution is mixed with a approximately equal amount of crystallization solution from a reservoir. The drop is then suspended and sealed over the well solution, which contains the precipitant solution at the target concentration (figure 3.2b). The difference in precipitant concentration between the drop and the well solution is the driving force which causes water to evaporate from the drop until the concentration of the precipitant in the drop equals that of the well solution. Since the volume of the well solution is much larger than that of the drop (1 − 3 ml as compared to 1 − 20 μl), its dilution by water vapour leaving the drop is insignificant. During a vapour-diffusion experiment, the protein will start to concentrate from an unsaturated state (point A, concentration C_{pi} (Fig. 3.3b)) to reach a supersaturated state (point B). As the first crystals appear the concentration of protein will decrease (figure

3.3b). The crystal will then grow until the concentration of the protein in the drop reaches the solubility curve (point C, at concentration C_{pf}).

The rate of vapour-diffusion can also be controlled. It greatly depends on the surface area of the drop. By reducing the surface area one can slow the equilibration process dramatically. This can be achieved by means of the so-called "sandwich drop" technique (see figure 3.2b). The volume of a hanging drop is limited to about $8 - 10$ µl, since bigger volumes might cause the drop to fall off of the cover slip. This limitation is overcome in the "sitting drop" technique (figure 3.2b).

Dialysis techniques utilise diffusion and equilibration of small precipitant molecules through a semipermeable membrane. The protein solution is contained within the dialysis membrane which permits water and some precipitants to exchange while retaining the protein in the dialysis cell (figure 3.2c). The protein solution at the start of the dialysis experiment is in an undersaturated state (figure 3.3c). The concentration of the precipitant agent slowly changes as its diffusion through the membrane takes place. Thus, the system goes from an undersaturated state into the metastable region through the point **S** on the solubility curve. The advantage of dialysis over other methods is in the ease with which the precipitating solution can be varied, simply by moving the entire dialysis button or sack from one condition to another. Thus, the protein solution can be continuously recycled until the correct conditions for crystallisation are found. The technique allows for salting in and salting out of the sample, as well as using a pH change to induce crystallization. For example, a sample requiring a high ionic strength for solubility can be dialyzed against a solution of low ionic strength to salt it out.

3.4 Crystal Types for EM

Efforts to crystallize membrane proteins for EM studies may result in production of different crystal types: flat single layers, tubular, vesicular and flat multilayered crystals (Fig 3.4) (*50*). Up to now, high resolution protein structures were only determined from planar 2D crystals. However, a recent report (*51*) on the determination of the 3D structure

of the acetylcholine receptor at 4.0 Å from tubular crystals has paved a way to use protein tubular crystals in structural studies.

3.4.1 2D Sheets

This kind of crystals (Fig. 3.4a) has lead to high resolution structure determination of various proteins such as bacteriorhodopsin, light harvesting complex II, tubulin, aquaporin and glutatione transferase (*32-36*). There are two different types of 2D crystals. Only membrane proteins form type I crystals (Fig. 3.5). Type II crystals are mostly for soluble proteins and macromolecule complexes because a regular layers of hydrocarbon based lipids is not stable in the presence of detergents needed to keep membrane proteins in solution (Fig. 3.5). Since planar 2D crystals can grow up to 30 μm in diameter, the combination of diffraction experiments with imaging is possible.

The main difficulty of EC is to obtain large single layer 2D crystals. The most possible reasons which diminish crystal and data quality are low mechanical stability, low specimen stability against beam damage, difficulties to record images of highly-tilted crystals and surface bending.
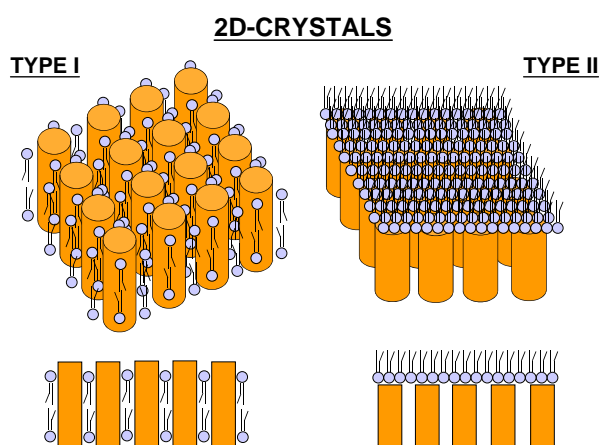
Figure 3.5: Types of 2D crystals.
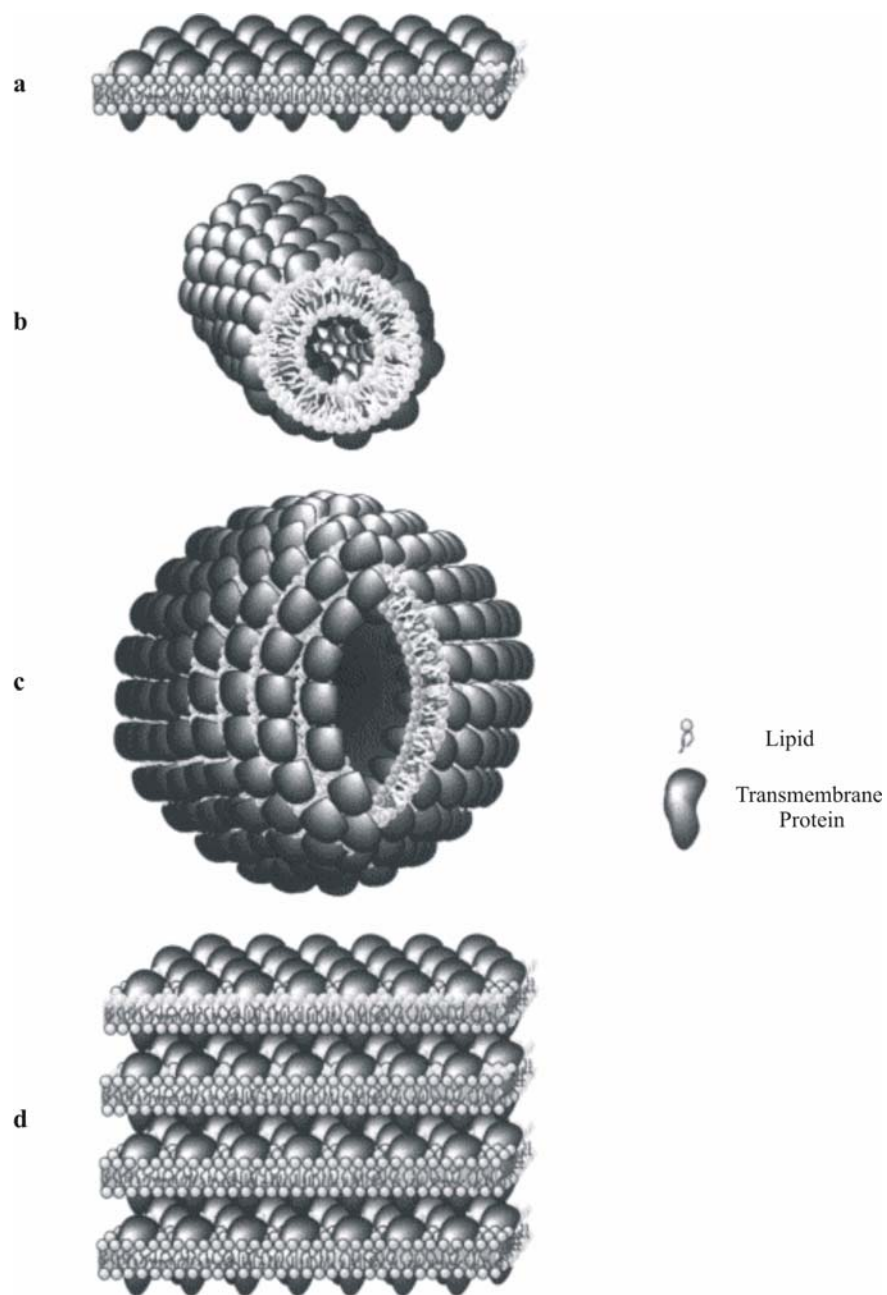
Figure 3.4: Four different crystal types. Crystallization experiments gives rise to four major types of crystals: a) sheets; b) tubes; c) vesicular and d) multilayered. Adopted from *(50).*

3.4.2 Tubes

As mentioned above, only one study was successful to obtain high resolution data from tubular crystals (*51*) (Fig. 3.4b). Since such tubes showed helical symmetry, a single image of a tube contains different projection orientations of the molecule along the tube. It means that no tilting of the sample is required. However, one of the major drawbacks lay in the variability of the tube's helical periodicity. Although Miyazawa *et al.* (*51*) have shown that tubes could be analyzed as short segments rather than in their full length, no other structure using tubular crystals has been solved. This method is still under development and it is not clear whether tubular crystals will be a general alternative to sheets.

3.4.3 Vesicles

The vesicular crystals are the easiest to obtain (Fig. 3.4c). They can be compared to an empty balloon whose cloth would be lipid bilayers with reconstituted membrane proteins (*50*). Even though vesicles are easily formed during crystallization experiments, they usually retain only a short-range order of the lattice. This is caused by the collapse of the spherical specimen during preparation of the EM sample and because of the intrinsic curvature of the vesicles. At the present time, this kind of crystals is not used for reconstruction.

3.4.4 Multilayered Crystals

Multilayered crystals (Fig. 3.4d) can be of two types:

1. Stack of singles layers which may (or may not) be in register;
2. Thin 3D crystals which have quasi-infinite periodicity in x,y directions and limited periodicity in the third direction.

Multilayered crystals represent an intermediate step between single layer crystals, i.e. 2D, and real 3D crystals. This kind of crystals can be obtained for soluble and membrane proteins as well. The main advantages of the multilayered crystals are the much stronger specimen stability due to additional interactions between layers and higher crystal order. Since a larger number of unit cells contributes to the diffraction signal of a given crystal area (if all layers are in register), the diffraction intensities are much stronger than in the case of single layer crystals.

3.5 Protein Crystallization Approaches for EM

Production of 2D crystals is the truly limiting factor of the structure determination process. Since natural ordering of proteins as 2D crystals is rare, different methods of forming 2D arrays were developed to expand the range of the electron crystallographic approach (Table 3.1):

1. **The *in situ* approach.** The principle of the *in situ* approach is to raise the concentration of the protein above the critical threshold required for lattice formation by extracting an excess of bulk lipid from the membranes. Another way of looking at this technique is that the removal of excess lipid induces a phase separation that segregates the protein into a crystalline domain. There are no restrictions as to what detergents are suitable provided the detergent does not solubilize the membrane protein of interest under the conditions being used. The *in situ* approach is feasible if the target protein constitutes the major fraction of the total membrane protein in a particular membrane fraction.

2. **The reconstitution approach.** The reconstitution technique is the traditional approach to 2D-crystallization and, to date, has produced the largest number of crystalline specimen. Indeed, any membrane protein which can be expressed and purified can be reconstituted in a lipid membrane by
   - addition of small additives**;**
   - phospholipase A_2 treatment to reduce the quantity of lipid in the membrane;

- temperature transition;

- extraction of other membrane proteins by the use of detergents.

All reconstitution methods start with purified proteins and lipids (native or synthetic lipids). The detergents concentration is then lowered in order to reconstitute an artificial membrane with the protein crystallized.

3. **The surface crystallization approach** (for membrane proteins). The surface crystallization technique exploits the tendency of membrane proteins to spontaneously form two-dimensional arrays at the interface between two phases such as the air-water or carbon-water. The latter can be exploited to grow 2D-crystals directly on the carbon support film of the electron microscope grid. This has been demonstrated for the plasma membrane H^+-ATPase, which was the first, and so far only, protein crystallized by this approach (*52*). A different way of looking at this approach is that it represents a substrate-induced 3D-crystallization, which is interrupted at the stage where the crystal is only 1-2 layers thick.

4. **Lipid monolayer crystallization** (for soluble proteins)**.** This approach exploits the enrichment of proteins at a lipid monolayer interface. The high protein concentration in vicinity of the lipid layer can induce the formation of 2D crystals. Several tricks have been used to "immobilize" the protein at the lipid-buffer interface. For instance, small amounts of charged lipids can be used to bind the protein by electrostatic interactions. More specific interactions can be enforced by using lipids with defined affinity-labels such as Ni-binding headgroups. Such "bait-lipids" can specifically bind to any histidine-tagged protein.

Table 3.1: Comparison of 2D crystallization techniques

Method	Advantages	Disadvantages
In situ approach	Least disruptive because protein is never removed from original membrane; Absence of crystallographic screw-axis makes time resolved studies easier because substrates can reach all target sites on the lattice; Absence of crystallographic screw-axis allows determination of the absolute orientation of the molecule within the membrane.	Size of the unit cell may vary between different preparations; Restricted number of crystallographic symmetries that are possible (p1, p2, p3, p4 and p6); Crystals may be small and (heavily) contaminated with non-crystalline membrane material.
Reconstitution approach	Offers the most control about crystallization conditions from all approaches. This may result in a higher degree of reproducibility; Allows crystallographic symmetries with screw-axis because molecules may insert into the membrane with both possible orientations; The ratio of crystalline to non-crystalline material may be higher; The crystals may be large enough to collect electron diffraction data.	Needs purified protein. Purification may remove trace lipids that are required for function or crystallization; Since reconstitution results in formation of vesicles or tubes, adding substrates for time resolved studies does not work well if the protein is arranged on a lattice that contains screw-axis.
Surface crystallization	May result in large arrays, suitable for electron diffraction; Arrays form within minutes to periods of a few hours; Crystals are formed directly on the grid; Needs very little protein.	Large crystals may not be coherent across large areas; The presence of precipitants such as PEGs in the crystallization solution causes high background in the diffraction patterns; The crystals are held together by detergent. Any changes to the environment are likely to disorder the crystals. This makes them unsuitable for time-resolved studies and observation in negative stain.
Lipid monolayer crystallization	May result in large arrays, suitable for electron diffraction; Arrays form within a few minutes to hours; Needs very little protein.	Large crystals may not be coherent across large areas; The crystals are extremely fragile, and transfer to the electron microscope grid is tricky.

3.6 Crystal Disorders

One unavoidable aspect of crystallizing a protein from scratch is the need for a large number of experiments exploring the various conditions that are necessary for successful crystal growth. The crystallisation of an uncharacterised –or even a well-characterised – macromolecule is not nearly as straightforward as it might seem. In fact, there are a vast number of possible conditions, which must be analysed for their ability to trigger the crystallisation process. In addition, the kinetics of nucleation and crystal growth varies in a random fashion. Thus, a condition, which is viable for crystal growth but unstable for nucleation, may take days to months to produce crystals, whereas another condition might produce crystals within a few minutes.

The protein crystallographer is primarily concerned with identifying one solution condition that will produce a relatively large (>100 μm in all dimensions) crystal that diffracts to high resolution. However, fundamental studies of protein solubility, nucleation and crystal growth usually are not performed. The reason for that is the lack of material (usually only a few milligrams of protein is available). At the present time, protein crystallization relies on screening techniques that use micromethods. Typical methods allow one to screen >100 solution conditions with just a few milligrams of protein. Most screening kids base on a sparse matrix approach, which uses a relatively limited number of precipitants. If initial microcrystals are produced during the screening trials, further optimisation is conducted to produce large crystals at the optimised condition.

Very often the step of moving from microcrystalline samples to large crystals is unsuccessful. Even when larger crystals are available, one can be faced with two major crystal disorders:

- **High mosaicity**. In a perfect crystal, all molecules form an ideal crystal lattice. According to Darwin's theory (*53*) an ideally perfect crystal does not occur in nature: real crystals have locally ideally perfect regions oriented in a random manner. This effect of random distribution is called "mosaicity". A high degree of mosaicity decreases the maximal resolution of the obtainable diffraction data

and "smears" the diffraction pattern. It is important to note that such crystals have only short-distance ordering of the lattice.

- During crystal growth nuclei may sometimes come together on crystal surfaces in entirely different orientations in ways that achieve good molecular packing but violate the crystal symmetry. This kind of disorder is called **twinning**.

In some cases, when twinning or mosaicity are the problem, small crystals can be of superior quality to larger ones. Unfortunately, macromolecule microcrystal cannot be used in conventional X-ray crystallography but may be usable for Electron Crystallography. Therefore, high resolution information about single protein molecules and protein-protein complexes can be obtained.

3.7 Conclusion

Theoretical aspects and various experimental methods of protein crystal growth are presented in this chapter. It is clear from the discussion above that the production of good, well-ordered crystals, which are suitable for EM or X-ray analysis form the bottleneck for structural studies. One possibility of relief is the use of multilayered crystals for EC. Often this crystal type seems to be a by-product of 3D- and 2D-crystallisation experiments.

Chapter 4

Data Collection

In this chapter we discuss all steps necessary for the recording and quantitative interpretation of diffraction patterns from multilayered catalase crystals. In addition, different methods for data processing, such as determination of the tilt geometry by a least-squares algorithm or by an analysis of Laue zones, and their application are presented.

4.1 Ewald Construction

When a crystal is placed into an incident X-ray or electron beam, only those reciprocal lattice points can be recorded which fulfil the Laue diffraction condition. One of the geometrical representations of this law is the Ewald construction of the scattering process (Figure 4.1).

The above considerations apply to all forms of crystallography, only the specific geometry can be different, i.e. the radius of the Ewald sphere varies. The basis of the geometrical construction is energy conservation which leads to an unchanged magnitude of the wavelength of elastically scattered X-rays or electrons.

All possible wave vectors $\vec{k}$ of the scattered wave lie on a sphere. According to the Laue criterion, a spot in the diffraction pattern is formed only when the scattering vector $\vec{S}$ is an integral multiple of the lattice vector in reciprocal space. In other words, the difference
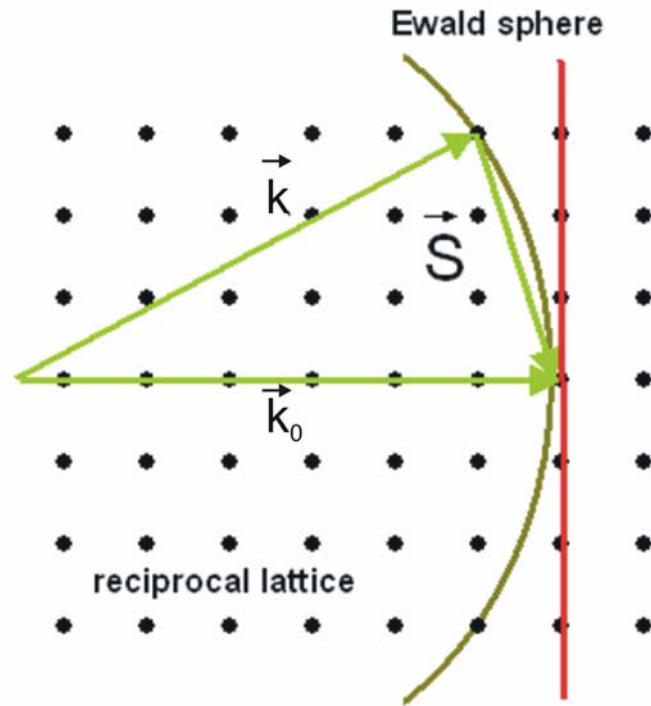
Figure 4.1: Ewald Construction. Depending on the radius, the Ewald sphere intersects a number of points in reciprocal space. This leads to diffraction maxima on the detection medium, as the Laue criterion for constructive interference is fulfilled. As the wavelength for electrons is much shorter than for X-rays, the Ewald sphere in the case of EC is flattened (red line).

$\vec{S} = \vec{k}_0 - \vec{k}$ must be equal to an integer-number multiple of the lattice vector to form a diffraction spot. This criterion is fulfilled in those cases where the Ewald sphere intersects a point of the reciprocal lattice as depicted in figure 4.1. The radius of this sphere is $\left|\vec{k}\right|$ which is equal to $\dfrac{2\pi}{\lambda}$. The shorter the wavelength, the larger the radius of the Ewald sphere.

In case of X-ray crystallography the Ewald sphere has a relatively small radius. In electron crystallography the radius of the Ewald sphere is quite large, as the wavelength of the electrons is of the order 0.04 Å compared to 1.54 Å for X-rays. This means that the sphere surface is flattened. One problem obvious from the geometrical Ewald construction is the fact, that not all spots in reciprocal space are sampled for a single tilt geometry. Therefore series of tilted diffraction patterns must be recorded in case of EC (comparable to the
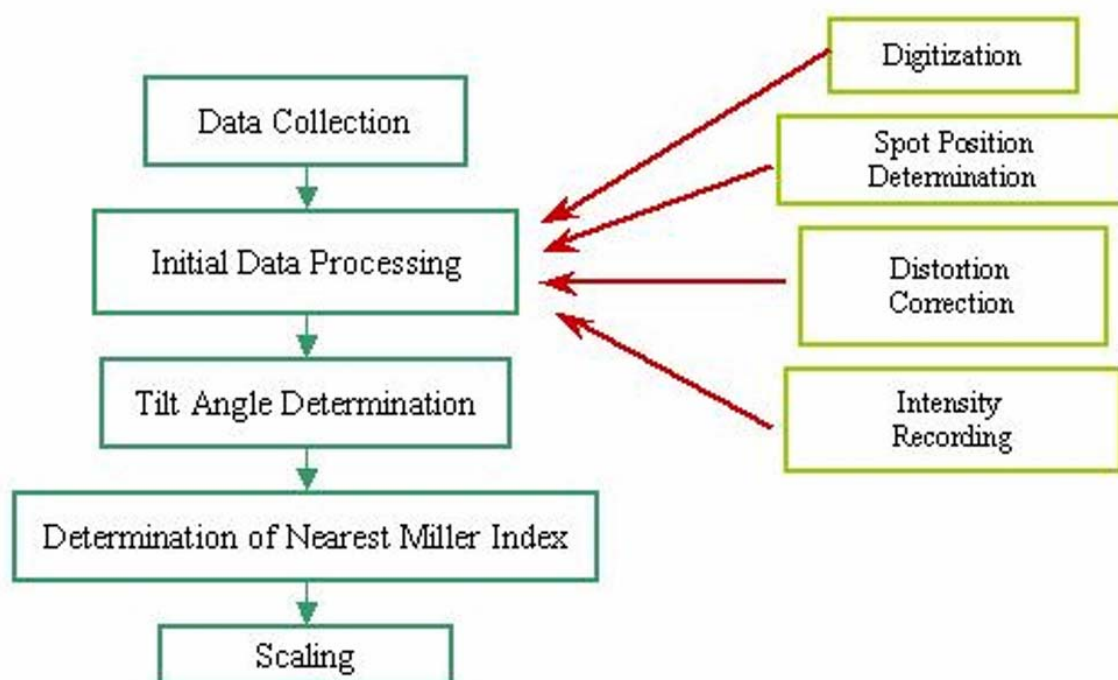
Figure 4.2: Flowchart showing a schematic overview of the data processing steps.

continuous rotation method in X-ray crystallography) to sample the entire 3D volume in reciprocal space.

4.2 Data Processing Strategy

The data processing strategy is depicted in figure 4.2. It includes the following steps:

- The diffraction patterns from crystals at different tilting geometry are collected in an EC experiment. This step includes visualization of the crystal on a carbon film support and recording diffraction data with the proper detection medium.

- These diffraction patterns, recorded on Imaging Plates or photo negatives, are processed digitally. The positions of the spots are measured and corrected for different imaging distortion effects with sub-pixel accuracy. Then the intensity of each spot has to be recorded (software package developed previously by Dimmeler and Vossen (*21;24*) and modified for the current studies).

- Using the positions of the spots, the tilt geometry of the crystal can be determined by the fitting of seven parameters, as described by Dimmeler and Schröder (*23*).

- When the tilt geometry is known, the nearest Miller index for each spot is determined.

- After the Miller indices have been assigned it is necessary to place each diffraction pattern on a common relative scale. This scaling is performed to remove effects of different exposure times, crystal size, and radiation damage.

4.3 The Study of Catalase as Test Specimen

Multilayered crystals of bovine liver catalase were used for electron diffraction experiments. Bovine liver catalase is an iron containing enzyme of MW 250000 Da which
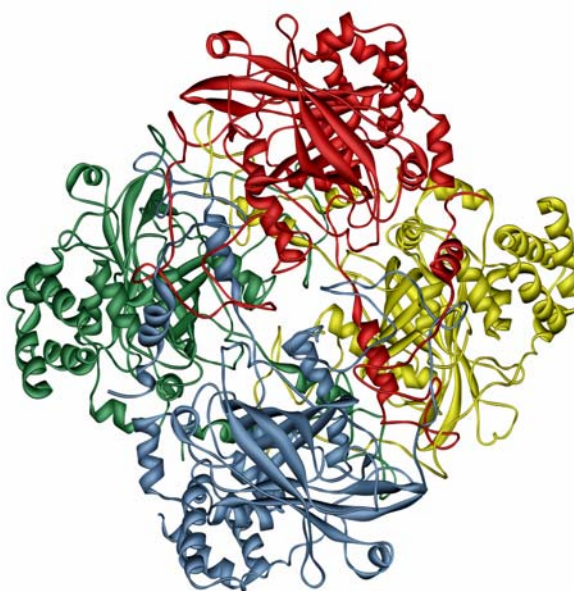
Figure 4.3: Bovine liver catalase tetramer solved by X-ray diffraction (PDB code – 4BLC). Each subunit is encoded by a different colour. The Figure was prepared using "WebLab Viewer".

catalyzes the reaction

$$H_2O_2 + H_2O_2 \rightarrow 2\,(H_2 + O_2\,).$$
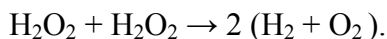
Catalysing this reaction, it protects cells from the toxic effects of hydrogen peroxide. The enzyme contains four identical subunits, which form a tetrameric complex (Figure 4.3). Crystals of bovine liver catalase are easily obtained, growing in different crystal forms (*54*). They have been intensively examined by electron crystallography using a number of techniques and by X-ray methods. In some crystallization conditions the enzyme forms big 3D crystals suitable for X-ray analysis. However, in others, the enzyme crystallizes as long needles or thin flat platelets. The latter form is extremely useful for electron microscopy studies and recording of electron diffraction patterns and was used in this work (for the exact crystallization condition see Appendix A). The thin 3D microcrystals have orthorhombic lattice parameters a=69 Å, b=173.5 Å, c=206 Å and a $P2_12_12_1$ space group. They were studied in the EM as frozen hydrated samples (*55*) and diffracted to a resolution of about 2.6 Å.

4.4 Localization and Choice of Crystals

After the sample has been prepared (for the details on sample preparation (quick freezing after blotting) see Appendix A) and transferred into the electron microscope, appropriate crystals embedded in an appropriate ice layer must be identified. Therefore, search for a good crystal is a two step procedure. First, the grid has to be screened for areas that have an adequately thin ice layer. The ice layer thickness can be judged from the energy loss spectrum. The second step is to find crystals that are of appropriate size and good quality. These steps are not trivial because the grid has to be screened at low electron dose and low magnification to reduce beam damage.

Furthermore, since proteins consist of light atoms, the contrast between a protein crystal and the embedding ice is extremely low. An increase of contrast can be achieved by using an aperture in the back focal plane of the objective lens, which leads to an increase in
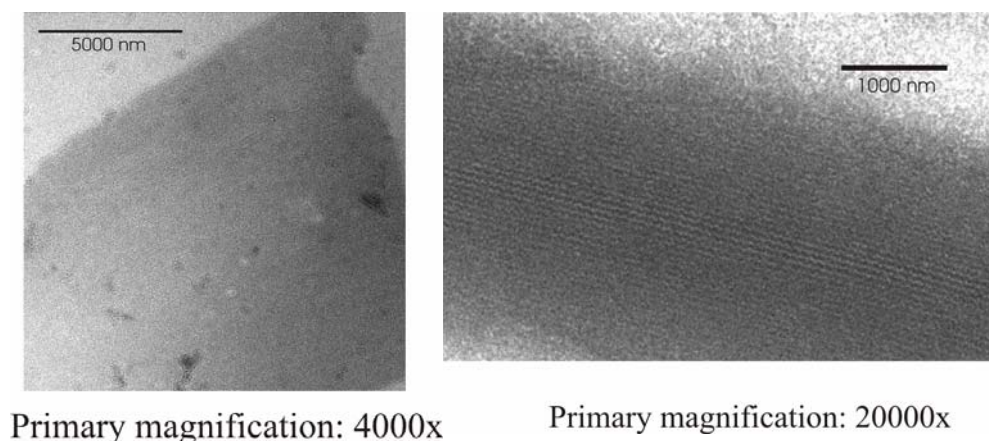
Figure 4.4: Electron micrographs of multilayered bovine liver catalase crystals in vitreous ice. The images clearly show the periodicity of the crystal lattice. Zero-loss energy filtered images were recorded using a LEO EFTEM 912 Ω microscope at an electron energy of 120 keV and a slow-scan CCD camera.

amplitude contrast. However, this aperture is limiting the resolution. All information at higher scattering angles (i.e. higher resolution) is cut off.

A better contrast improving technique is the use of an energy filter. In this work, zero-loss energy filtered images have been used to provide maximal contrast. By filtering out inelastically scattered electrons, the signal-to-noise ratio of the elastically scattered electrons is improved (*56*). The main advantage of zero-loss imaging in the case of multilayered crystals is an increased amplitude contrast of the crystal area (*57*) which allows an easy identification of the crystals. The increased mass thickness of protein crystals shows as an increased amplitude contrast, i.e. the crystals appear much darker than without energy filtering.

Crystals are usually defined by sharp edges that often have pointed corners (Fig. 4.4). The overall shape of the crystal depends on the crystal lattice type. For example, cubic lattice crystals often form crystals that have 90° corners, while hexagonal ones form 120°. These features can be used to find crystals on the EM grid at low magnification.

Once a candidate is located, the quality of diffraction data has to be tested by visualization of a diffraction pattern on the fluorescence screen under low dose conditions, i.e. at about 2-3% of the allowed total electron dose. This step is important because only about 10 % of

the catalase crystals are of good quality. Several parameters may influence the diffraction data quality:

- Crystal size. Experiments have shown that crystals must be at least 3-4 μm in diameter to record high resolution diffraction data. Only large crystals guarantee diffraction patterns that have a sufficient signal-to-noise ratio to allow quantitative data interpretation. The larger the crystal, the stronger the signal. However, the large crystals may have more layers in the Z-direction and will be "too thick" for an electron beam. Crystals of different size must be tested to find optimal conditions for recording of a high resolution diffraction pattern.

- Crystal quality. The crystal must have high periodicity. If the order is not very high, diffraction patterns will not show high resolution data. Sometimes the catalase crystals are broken and therefore useless for electron diffraction studies. One reason could be that crystals are destroyed or broken up during sample transfer. This phenomenon is a typical experimental problem of sample preparation and depends mainly on the mechanical properties of the crystal.

- Beam damage. One problem of the high elastic scattering cross-section of electrons is the equally high – and in fact even higher – inelastic scattering cross-section. Inelastic scattering deposits energy in the sample causing the destruction of the sample. A possible mechanism of the damage is the intense ionization of the specimen, which results in the formation of free radicals and ions. This causes bond scission and thus formation of molecular fragments. Even at cryogenic temperature the free radicals (in particularly H·, ·OH and hydrated electrons) can undergo diffusion and produce cross-linking or further chain scissions. Damage of the secondary structure of proteins occurs at an electron dose of less than 1 $e^-/Å^2$, and further exposure causes the tertiary structure to undergo dramatic reorganization following the loss of specific groups and altered structural composition. The long exposure of the catalase crystals results in reduction of the maximal resolution. The maximal irradiation which can be deposit into the crystals is 10 $e^-/Å^2$. It should be noted that all testing steps (grid screening, preliminary diffraction) must be carried out at low electron dose to minimize radiation damage effects.

4.5 Electron Diffraction and Tilt Series

The last step in data collection is the recording of diffraction patterns from tilted crystals (Figure 4.5) to provide sufficient sampling of the Fourier space. Typically, 3-5 diffraction patterns were recorded from each crystal on Imaging Plates (IP). The advantages of IPs are detection linearity, high dynamic range and large recording area. The average electron dose for one pattern was 2.5 $e^-/Å^2$.

A crucial improvement in diffraction pattern quality was obtained by the use of an energy filter. Usually, the diffraction pattern includes a strong diffuse background, which reduces the signal-to-noise ratio and makes intensity recording more difficult. Since this background results primarily from inelastically scattered electrons, it can be significantly reduced by using an electron energy filter in zero-loss mode, which selects only elastically scattered electrons for data recording.

$\Theta= 3°$ (2nd pattern in tilt series)

$\Theta= 29°$ (1st pattern in tilt series)
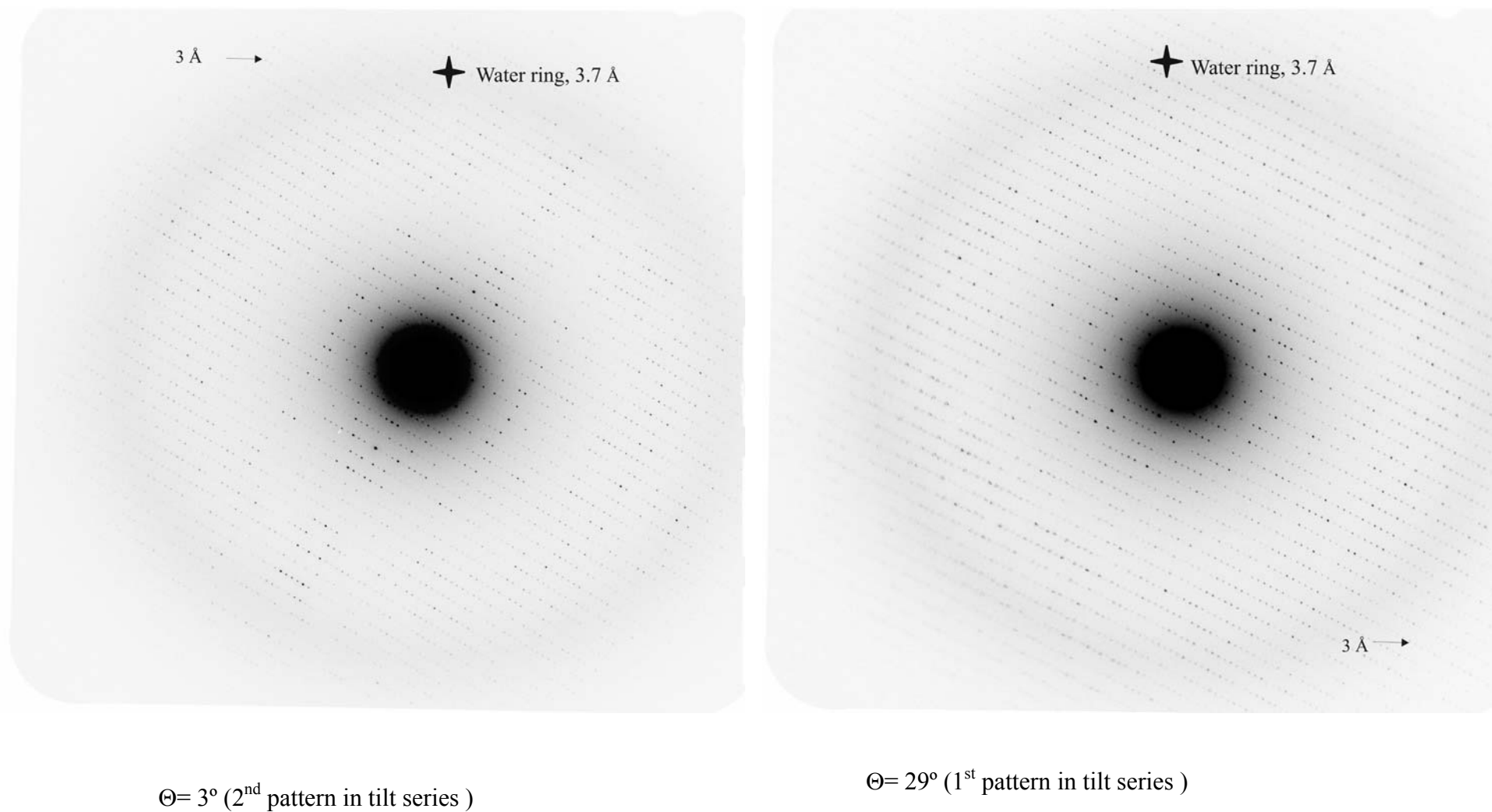
Figure 4.5: Catalase diffraction patterns recorded at different tilt angles. A resolution better than 3 Å is achievable (arrows).

4.6 Tilt Geometry and Necessary Completeness

One of the important parameters in crystallography is data completeness which is defined as the ratio between the number of observed reflections and the number of theoretically possible reflections. Any reflection not recorded and excluded from the reconstruction causes a loss of signal proportional to the amplitude of the reflection. In other words, a missing reflection is equivalent to a measurement with a 100% error for this particular reflection. To design a data collection protocol so that a data set is complete, i.e., all unique reflections are measured at least once, is not straightforward. Geometric considerations must be taken into account. Two main aspects such as optimal sampling of reciprocal space and minimization of the "missing cone effect" have to be considered.

4.6.1 Sampling of Fourier Space

The crystal orientation is described by six parameters. The six parameters are (Fig. 4.6) tilt angle Θ, tilt axis angle Φ (angle between tilt axis and unit cell axis a^* in reciprocal space), angle ε between the former untilted a^* axis (where the a^* axis was located before tilting) and the horizontal axis of the film pointing to the right, coordinates x_m and y_m of the centre of the tilted diffraction pattern on the film, and a calibration factor F_{angpix} between Å and pixel in the digitized pattern for the untilted geometry.

To record all reflections, Fourier space has to be sampled by variation of the angles Θ and Φ. However, only the angle Θ can be changed during an experiment by tilting the cryo-specimen holder, while the angle Φ is fixed, and cannot be controlled by the sample preparation. It depends on the orientation of the crystal on the carbon film and is completely random in its orientation relative to the tilt axis. It is important to note that the density of the (Θ, Φ) distribution must be isotropic to provide isotropic resolution of the reconstruction. The necessary increment of tilting depends on the unit cell parameters, but for protein crystals should be of the order of 1° or less. More precise sampling will give a higher redundancy of the data and better statistics, but the data collection will be more time

consuming. Furthermore, the increment depends on the desirable resolution. The higher the desired resolution, the smaller the increment must be.

Another aspect is the decrease of diffraction amplitudes at high resolution due to a combination of sample, experimental, mechanical, and processing defects[9] (*27*). This amplitude loss makes it necessary to collect 10-100 times more data than theoretically necessary to obtain statistically significant high resolution data . As shown in Fig. 4.7, diffraction patterns included in this study are almost randomly distributed in space. In this work, we used 55 diffraction patterns from 17 tilt series to calculate the final reconstruction (see Fig. 4.8).
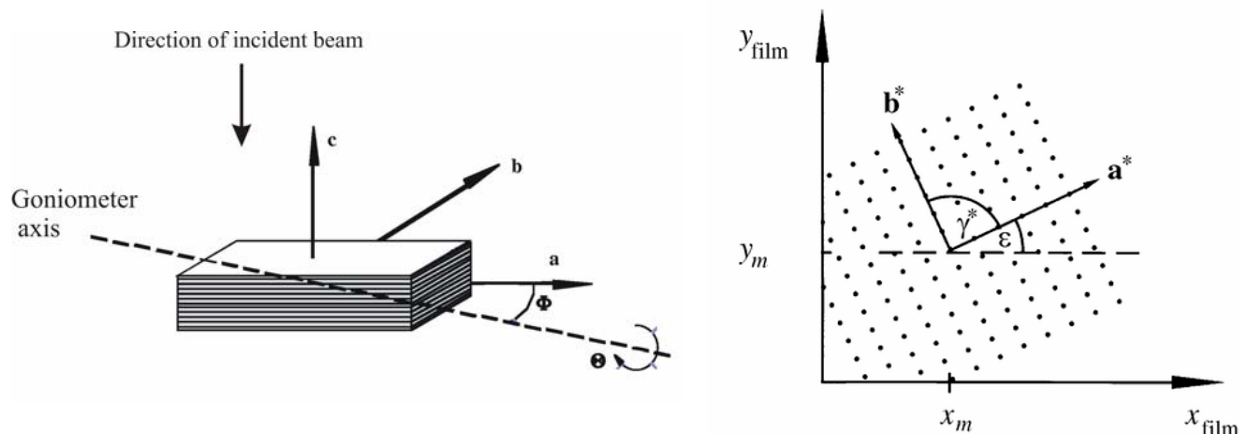
Figure 4.6: Tilt angle geometry. In general, seven parameters are needed to describe the position of a crystal in real space. The tilt angle Θ and tilt axis angle Φ describe the orientation of the Ewald intersection in reciprocal space. The parameters x_m , y_m and ε encode the position of the diffraction pattern in a digitized image. A further parameter F_{angpix} describes the scaling of the pattern in the pixel representation. For monoclinic crystals with the unique axis **c**, a further parameter γ^* is needed for the additional lattice angle.

---

[9] In the case of X-ray crystallography, structure factor amplitude decreases with resolution ($\sin\theta/\lambda$) due to vibration motion of atoms about their equilibrium positions. This is also true for electron diffraction. However, the additional effects mentioned above contribute also severely.
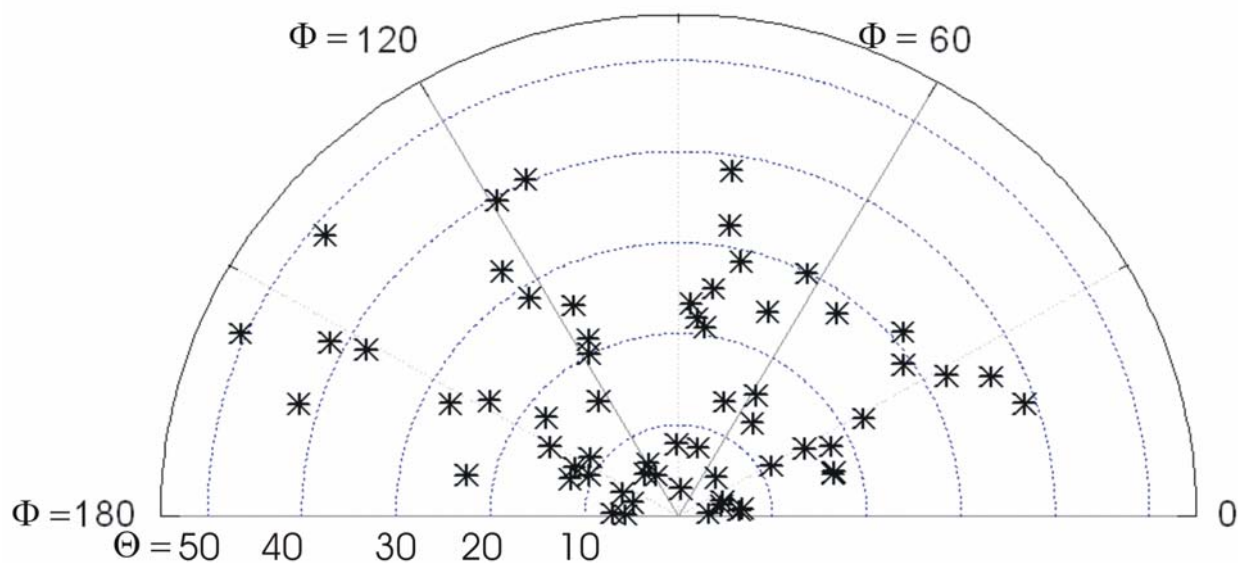
Figure 4.7: Polar (Θ, Φ) distribution of diffraction patterns included in this study. Θ and Φ describe the angular offset of the unit cell axis in respect to the tilting axis.
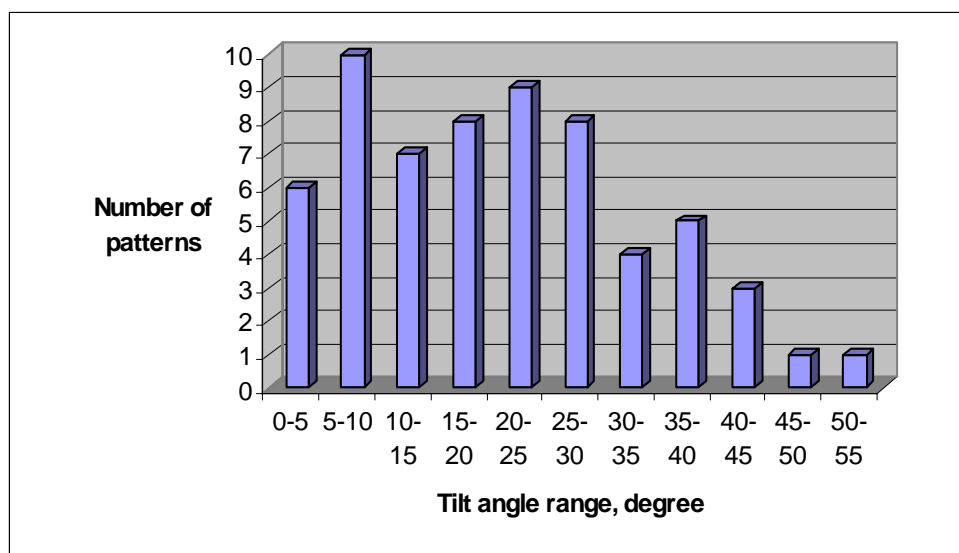
Figure 4.8: The diagram shows a distribution of diffraction patterns according to tilt angle Θ.

4.6.2 Missing Cone Effect

The significant limitation that reduces data completeness is the so-called "missing cone effect". In practice, it is very difficult to collect diffraction data from higher tilt angles, and thus a cone-shaped region along the axis of the incident electron beam in the reciprocal space is not sampled. The range of possible tilt angles in most microscopes is restricted to less than 60°. This results in a missing cone of diffraction data which reduces the resolution of structural features in the direction perpendicular to the plane of the specimen.

How much completeness do we need? To refine an electron density map at atomic resolution with small details, i.e., conformations of side chains, the data completeness should be at least 70-80%[10]. Lower completeness results in noisy electron density maps which are hardly interpretable. Although a full dataset cannot be collected with a modern
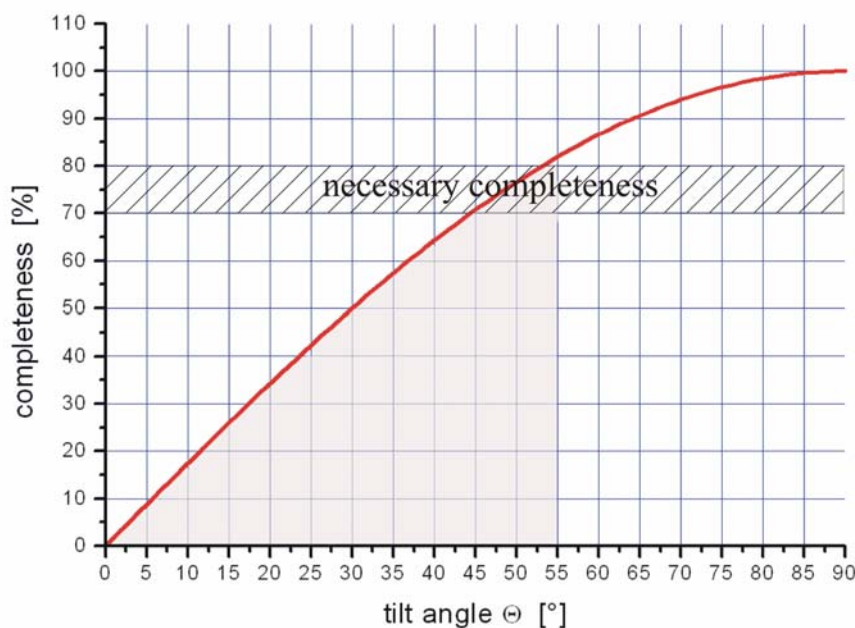
Figure 4.9: Maximal possible completeness for different tilt angles in space group P1.

---

[10] Numerical simulations were performed using 12 datasets with different completeness (see Chapter 6.7) to analyze an effect of incomplete (isotopic and anisotropic) data on reconstructed density interpretation. It was shown (data not presented) that data with completeness of 70-80 % (both isotropic and anisotropic) is enough to built a molecular model according to the calculated density.

cryoholder, it is still possible to record diffraction patterns at tilt angels up to ±55°-60°: which corresponds to 80-85 % completeness even for space group P1 (Figure 4.9). Higher symmetry would reduce the required volume in Fourier space. For example, for the cubic space group I432 or F432, the necessary tilt angle range would be only about 35° to collect complete data[11].

We collected a diffraction dataset from multilayered catalase crystals with a completeness of 50-70% depending on the resolution cut-off (Table 4.1, Fig. 4.10). Since the goal of this thesis is to demonstrate the proof of principle of using multilayered protein crystals for a structural analysis, we did not continue to collect the maximal theoretically possible completeness. The bovine catalase structure is known at high resolution (*58*) and there is no scientific interest from the biological point of view to solve the same structure in a different crystal form. There are no theoretical or experimental limitations, however, collection of a 80-85% complete dataset would need an enormous number of additional patterns.

Table 4.1: Data completeness

Resolution, [Å]	Number of observed unique spots	Number of theoretically possible spots	Completeness, [%]
20 – 19	20	31	65
19 – 18	28	38	74
18 – 17	29	47	59
17 – 16	49	69	71
16 – 15	56	82	69
15 – 14	65	100	65
14 – 13	93	139	67
13 – 12	122	187	65
12 – 11	166	251	66
11 – 10	223	353	63
10 – 9	341	540	63
9 – 8	499	833	60
8 – 7	474	1351	58
7 – 6	1231	2388	52
6 – 5	2163	4636	47
5 – 4	4065	10368	40

---

[11] However, protein crystals are relativly rare observed in these spacegroups. According to an analysis of spacegroup frequencies of PDB holdings (Feb. 2000, 9481 entries) only 0.26% (0.25 %) proteins were crystallized in F432 (I432).
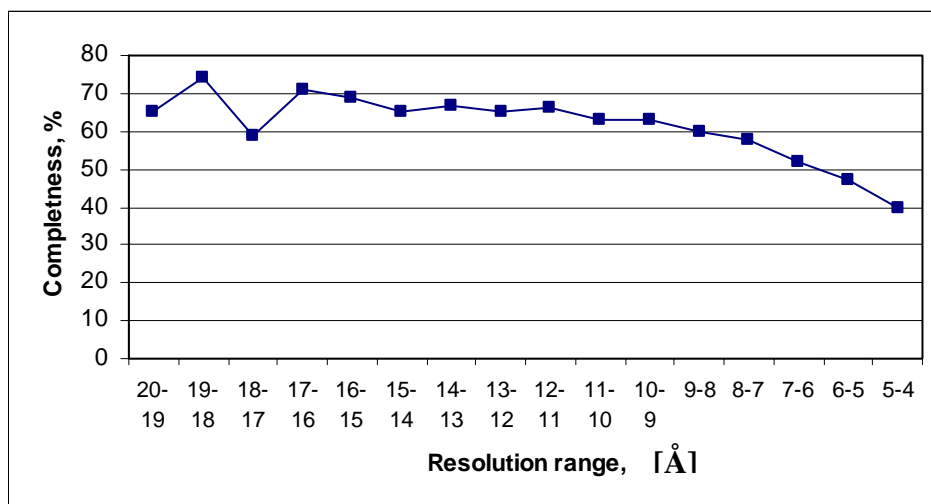
Figure 4.10: Diffraction data completeness statistics.

4.7 Tilt Angle Determination

4.7.1 Lattice Transform Profile of 3D, 2D and Multilayered Crystals

As it discussed in Chapter 2 and illustrated in Fig. 4.11, the sampling of Fourier data from thin, 3D crystals along $c*$ is not discrete and must be explicitly taken into account. One can distinguish three cases:

1) Diffraction from real 3D crystals produces a 3D lattice of discrete, well localized reflections (Fig. 4.11c).

2) Diffraction from 2D crystals produces a 2D array of lattice lines with continuous Fourier amplitude (Fig. 4.11a).

3) Thin, platelet-like 3D crystals fall in between these extremes and produce lattice lines comprising a series of Bragg reflections that are convoluted with sinc functions (Fig. 4.11b). The shape of the lattice can be calculated and the intensity can be found by formula 2.19.
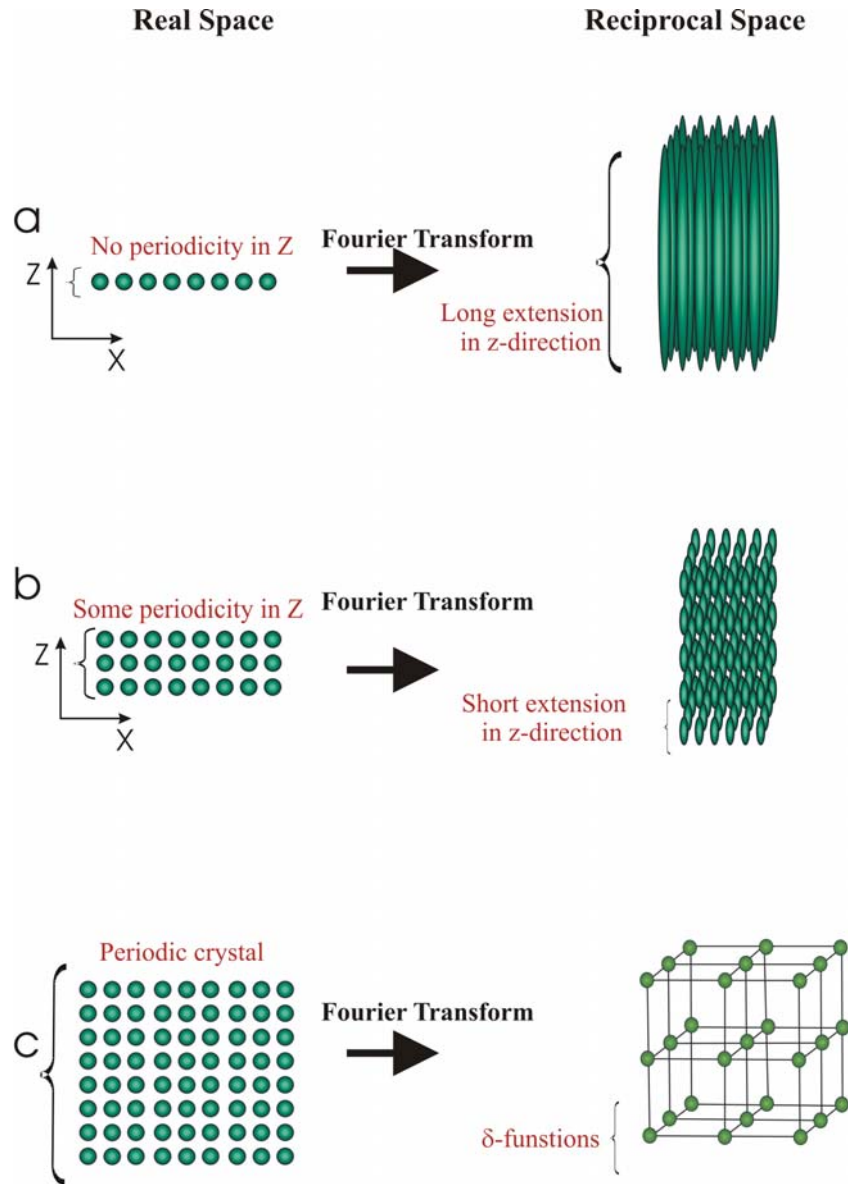
Figure 4.11: Effect of periodicity on density distribution of a) 2D; b) multilayered and c) true 3D crystals.

Dimmeler *et al*. (*22*) developed a method to determine the lattice-transformed density profiles from multilayered 3D crystals. However, in our studies the information has not been used. The strategy of the work was to adopt existent software for X-ray analysis, i.e. CCP4 (*59*), CNS (*60*), for the needs of EC. Implementation of the new lattice transform which is described by the Gaussian instead of the lattice transform of a real 3D crystal (array of δ-functions) would require rewriting the program codes. We found that the increased inelastic background at high tilt needs to be filtered out more efficiently, as it

adds a systematic error (*22*) to spot positions. With the current instrument (LEO 912 at 120 keV electron energy) data quality is limited. Nevertheless, later results show (Chapter 6) that using the lattice transform of a real 3D crystal the phase problem can be solved successfully. In the future, it will be desirable to use data from a better energy-filtered transmission electron microscope (EFTEM) (e.g. the now available LEO922) and to use Dimmeler's algorithm to increase the quality of the reconstructed structure.

4.7.2 Global Least-squares Method

The exact orientation of the crystal is extremely important for all structural studies since many patterns from different tilt angles have to be merged to obtain the final 3D dataset. It has been shown that an incorrect assignment of the tilt angles to the recorded diffraction intensities reduces the resolution of the final structure (*61;62*). As shown in Fig. 4.6 the crystal position and consequently the positions of the spots in the diffraction pattern is given by the geometry of the experiment (tilt angle Θ and tilt axis Φ). Since a typical diffraction pattern consists of a few thousand well-defined spots (see Fig. 4.5), the tilt parameters can be calculated using these spot positions. The elongated form of the reflections in the direction with limited periodicity leads to shifts in the diffraction pattern, denoted as the "characteristic shift" (*22*). This change in the position of the spots (compared to the spot positions expected from a 2D crystal) must be calculated. The characteristic shift can thus be seen as a random "noise" which cancels out on average because the sum of all shifts is the null vector. As a few thousand of spots are averaged, statistically, the characteristic shift has no effect on the tilt angle determination. Using the least-squares algorithm developed by Dimmeler *et al.* (*23*), the tilt angle Θ, tilt axis Φ, coordinates of the pattern center, rotation angle ε, the scaling factor F_{angpix} and a fit quality parameter χ^2 can be determined with very high accuracy from a single diffraction pattern.

In theory only one single pattern is needed to determine the tilt angle geometry. However, typically, several diffraction patterns at different tilt angles are recorded from one single crystal. It accelerates the data collection procedure and provides an additional crosscheck

for the calculated tilt angles as the experimental tilt angle increment is known. Table 4.2 shows the results of the least-squares fit for a typical tilt series of multilayered catalase crystal. This table shows that the fitted tilt angles $\Theta_{calculated}$ agree well with the coarse readings of the goniometer $\Theta_{experimental}$. However, one may note that, for example, for the tilt series #7, the difference between $\Theta_{experimental}$ and $\Theta_{calculated}$ is about 3 degrees. A possible reason for this is mechanical distortion (for example, bending) of the carbon film. In this case, the values of $\Delta\Theta_{experimental} = \Theta_{experimental, i} - \Theta_{experimental, j}$ and $\Delta\Theta_{calculated} = \Theta_{calculated, i} - \Theta_{calculated, j}$ (where i, j are numbers of the diffraction patterns in the tilt series) for diffraction patterns in the given tilt series must be calculated and compared to prove the accuracy of the tilt angle determination procedure. In the ideal case, $\Delta\Theta_{experimental}$ must be equal $\Delta\Theta_{calculated}$.

The error in the reciprocal coordinates z^* and consequently in indexing of the diffraction spot is proportional to $\cos\Theta$ (*23*). In other words, small errors in z^* lead to significant errors in the tilt angle determination. As a result of this, small tilt angles Θ ($\Theta < 10°$) cannot be determined with high accuracy. For such patterns we employed another method for the determination tilt geometry, which uses an analysis of Laue zones.

Table 4.2: Tilt series statistic.

Tilt series #	Diffraction pattern	Θ_{exp}	Θ_{calc} (least-squares algorithm)	Θ_{calc} (Analysis of Laue zones)	Φ	χ^2
1 (05_02_2003 xtal1)	Diff1	0	9.87	4.7	70.6	82.5
	Diff2	35	36.5	-	-28.0	70.2
	Diff3	-45	-41.6	-	-27.6	46.6
2 (05_02_2003 xtal3)	Diff1	39	39.26	-	113.7	17.8
	Diff2	21	21.56	-	116.1	31.9
	Diff3	-4	-9.3	5.6	-159	37.2
3 (05_02_2003 xtal5)	Diff1	21	27.1		152.8	13.9
	Diff2	6	12.2	-	151.8	6.66
	Diff3	-23	-17.7	-	142.4	2.83
4 (06_02_2003 xtal2)	Diff1	34	28.9	-	123.4	14.3
	Diff2	5	9.8	3.2	37.5	29.3
	Diff3	-16	-20.1	-	118.0	7.5
	Diff5	-36	-39.6	-	119.0	13
5 (06_02_2003 xtal3)	Diff1	0	9	10.8	2.36	36.4
	Diff2	27	25.2	-	83.6	23.8
	Diff3	-21	-20.9	-	82.2	54.9
6 (06_02_2003 xtal6)	Diff1	0	-10.28	5.8	-18	28.4
	Diff2	32	29.9	-	62.8	20
	Diff3	-48	-50.5	-	56.7	26
7 (13_02_2003 xtal1)	Diff1	0	-9.2	-5.0	46	11.3
	Diff2	14	11.3	-	145.5	10
	Diff3	27	23.7	-	147.7	6.4
8 (13_02_2003 xtal5)	Diff1	17	15.6	-	56.0	13.5
	Diff2	-14	-15.3	-	28.8	45.6
	Diff3	-31	-31.3	-	40.2	23.6
	Diff4	-4	-10.7	-6.5	5.0	67.8
9 (28_02_2003 xtal1)	Diff1	0	6.7	5.0	163	21
	Diff2	16	21.8	-	84.7	11.9
	Diff3	-14	-12.1	-11.3	87.5	15.2

4.7.3 Analysis of Laue Zones

Additional verification of the tilt geometry can be obtained from the presence of Laue zones near the (h, k , 0) projection. Laue zones are bands of reflections where the Ewald sphere intersects successive layers in the 3D reciprocal lattice. In figure 4.12a, the oblique lines correspond to surfaces of the Ewald sphere at three discrete tilt angles and the shaded boxes represent the intersection of these spheres with a 3D reciprocal lattice of a multilayered crystal. The shaded boxes produce bands of reflections and at higher tilt angles the zones are closer together and narrower until they eventually produce apparently continuous reflections (figure 4.12a, blue line). Thus, Laue zones are most prominent near the projection of the a*-b* plane (h,k,0).

These Laue zones are present in all diffraction patterns with $\Theta < 10°$. One of the examples (figure 4.12b) is the diffraction pattern from the tilt series #9. The Laue zones can be recognized easily as variations in intensities of the diffraction spots in the horizontal direction. According to our geometrical analysis (see Appendix A), the diffraction pattern is at a tilt angle of $\Theta = 5.0°$. The tilt angles determined by a Laue zone analysis for others patterns are summarized in Table 4.2 (fifth row). Comparing the accuracy of the least-squares algorithm and the analysis of Laue zones for low tilted diffraction patterns has shown that the latter method is more precise and obtains better agreement with values of the goniometer $\Theta_{experimental}$.

An additional crosscheck to verify the tilt geometry at low tilt angles is an analysis of locations of systematic absences. According to the diffraction theory, special reflections, so-called *systematic absences* or *extinctions*, have zero intensity as result of the symmetry of the structure of the unit cell. For example, the catalase crystals, which belong to the space group $P2_12_12_1$, the condition for the presence of systematic absences along a line 0k0 is k=2n. It means that every second reflection along this line must have zero intensity. This fact may be used to verify the tilt geometry for low tilted diffraction patterns. Figure 4.13 shows an example of such a pattern. At low resolution every second spot is missing (red arrows). However, since the Ewald sphere intersects the spots with $l \neq 0$ at high resolution, systematic absences are no longer observed.
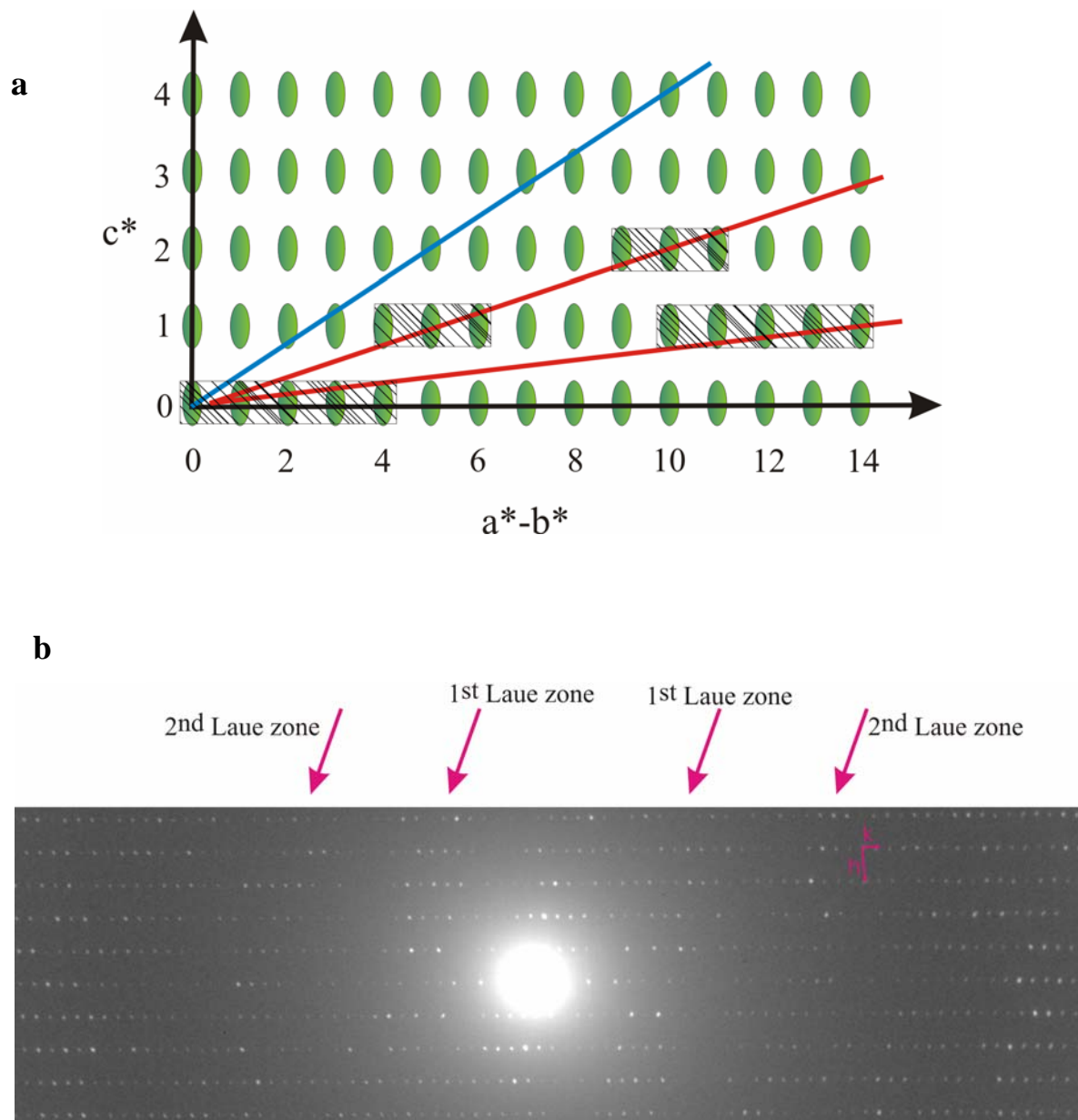
Figure 4.12: Analysis of Laue zones a) Schematic representation of the Laue zone concept. The shadowed boxes reproduce bands of reflection at low tilt angles (red lines); b) The diffraction pattern ($\Theta=5$) clearly shows the repeated modulation of intensities — the Laue zones.
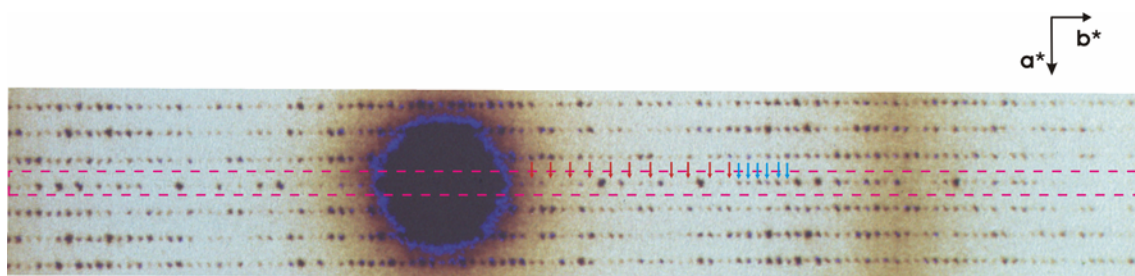
Figure 4.13: Appearance of systematic absences at low tilted diffraction patterns. Every second reflection is missing at the low resolution area (red arrows) along the 0kl line (red box) when l=0. However, at high resolution the Ewald sphere hits density profiles with l≠0 and, therefore, the observed spots do not fulfil the conditions for the presence of systematic absences.

4.8 Electron Diffraction Data Quality

The quality or the information content of a crystallographically determined 3D structure can only be as good as the quality of the underlying diffraction data. Therefore, the obtained diffraction data have to be carefully analyzed. The following criteria are available for assessing the quality of the raw crystallographic data:

1) **Redundancy**. Redundancy is defined as the number of independent observation per unique reflection in the final merged and symmetry-reduced dataset. Since electron diffraction data collection is to some extent influenced by counting statistic, the average measurement should become more accurate as more individual measurements are included. Therefore, a highly redundant dataset will normally be of higher quality than a dataset in which every reflection has only been measured once. Table 4.3 shows that redundancy N of the electron diffraction dataset varies from 3.8 at high resolution (4 Å) to 14.0 at low resolution (15 Å).

Table 4.3: Statistic of recorded diffraction spots

Resolution, [Å]	Number of reflections		Redundancy, N	R_{merge}, %
	observed	unique		
15	2542	182	14.0	28.8
10	7154	669	10.7	30.5
9	2991	340	8.8	27.9
8	3849	499	7.7	27.4
7	4857	782	6.2	27.4
6	6316	1231	5.1	25.7
5	9574	2163	4.4	27.5
4	15630	4065	3.8	26.9

It is obvious, that (higher) resolution and (lower) redundancy are correlated. This observation may be explained by two possible reasons:

1) High resolution spots are observed only in the first diffraction pattern of a tilt series (Chapter 4.4) and are not present in subsequent patterns due to radiation damage of the sample. Therefore, redundancy for lower resolution data will be higher when data from differently tilted diffraction patterns are merged together.

2) To sample the whole Fourier space, diffraction data must be recorded at different tilt geometries. One of the parameters, which describes the tilt geometry, is the tilt angle Θ (Chapter 4.6.1). Figure 4.14 shows the intersection of the density extension profile for multilayered crystals in reciprocal space at different tilt angles by the Ewald sphere. If θ is slightly changed, the Ewald sphere intersects spots with different *hkl* at the high resolution area, while low resolution spots will be measured twice.

2) **Signal strength**. The average strength or significance of the observed intensities is often expressed as the ratio of the intensity to the error of the intensity, i.e. I/σ. From a statistical point of view, I/σ is a good criterion for two reasons. Firstly, it defines a resolution
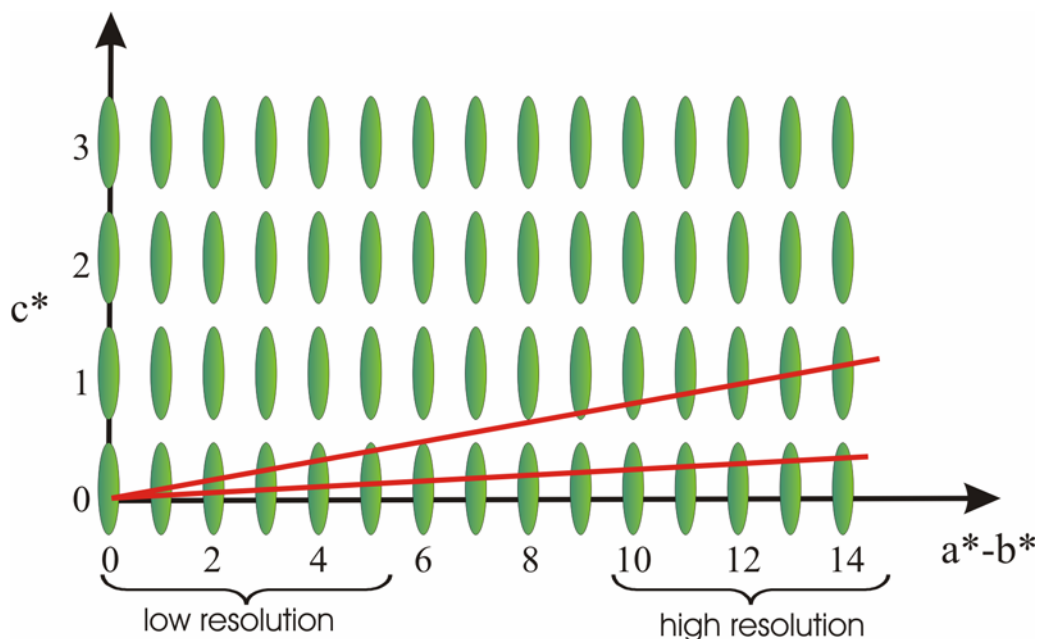
Figure 4.14: Intersection of the elongated density profiles by the Ewald sphere at different tilt angles Θ. If Θ changes slightly between two diffraction patterns, the low resolution spots are recorded twice.

"limit"[12] of the available data. Secondly, since σ is assigned to each reflection, its value can be used as weight, or "importance" of the given spot, for scaling different datasets.

Unfortunately, we cannot judge the quality of EM data by this parameter because an algorithm for I/σ calculation has not been implemented to the program find_intens (see Appendix A). In the future, it would be desirable to use signal strength statistics in data processing.

3) **Merging R values.** One of the commonly used indicators of the quality of diffraction data is the so-called merging R-factor (R_{merge}) which is defined as

$$R_{merge} = \frac{\sum_{h}\sum_{i}\left|I_{h,i}-\langle I_{h}\rangle\right|}{\sum_{h}\sum_{i}I_{h,i}} \qquad (4.1)$$

---

[12] For example, in X-ray crystallography diffraction data with $I/\sigma<3$ is not included in data processing.

where the outer sum (h) is over the unique reflections and the inner sum (i) is over the set of independent observations of each unique reflection (*63*). This statistic is supposed to reflect the spread of multiple observations of the unique reflections. These multiple observations may derive from symmetry-related reflections or from Friedel pairs. Unfortunately, R_{merge} is of relatively poor statistical value because its value increases with increasing redundancy. Low redundancy will always yield a lower R_{merge}, but at the same time result in less accurate data.

This dependence of R_{merge} on the redundancy (N) can be remedied by the redundancy-independent merging R-factor $R_{r.i.m.}$ (*64*), or R_{meas} (*65*):

$$R_{meas} = \frac{\sum_h [N/(N-1)]^{1/2} \sum_i \left| I_{h,i} - \langle I_h \rangle \right|}{\sum_h \sum_i I_{h,i}} \qquad (4.2)$$

It describes the precision of the individual measurements, independent of how often a given reflection has been measured, and may be used as a substitute for the conventional R_{merge} factor.

Table 4.3 shows values of R_{meas} for different resolution shells of the electron diffraction dataset. As could be noted, R_{meas} is about 25-30 % and does not strongly depend on resolution. These observations are in contrast to the usual statistic for X-ray diffraction data, where R_{meas} value is only a few percent for low resolution data and increases up to 20-30 % for high resolution.

The following effects may result in the poor statistics as observed in the case of EM data:

1) Dynamic scattering. The effect of dynamical scattering on diffraction data statistics for multilayered protein crystals was analyzed in detail by Vossen (*21*). He came to the conclusion that dynamic scattering breaks a Friedel symmetry thereby increasing R_{symm}[13]. The factor R_{symm} is a function of the thickness of the sample, electron energy used and resolution limit of the diffraction data. The higher the

---

[13] The factor R_{symm} describes the average deviation of the spots that comprise a Friedel pair and symmetry related reflections.

resolution that is included in the analysis, the higher the value of the factor R_{symm}. For example, simulations show that R_{symm} reaches 27% already for an one layer dataset using a bacteriorhodopsin simulation phantom at an electron energy of 200 keV (*21*).

2) Partial reflections. One serious problem of scaling diffraction data from differently tilted crystals is the merging of partial reflections. Although in our approximation the h,k,l indexes describe the spot in reciprocal space, the actual recorded intensities result from intersection of the Ewald sphere with different parts of the Gaussian reflection profile (Fig.4.15). Ideally, the total intensity distribution must be integrated by summing up the partial reflections. This integration algorithm was developed by Dimmeler (*22*). Unfortunately, the quality of the diffraction patterns used in our studies was not high enough to determine lattice-transform density profiles of multilayered catalase crystals and, therefore, to calculate an integral intensity.
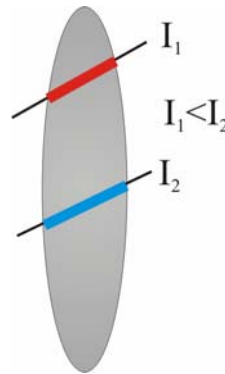
Figure 4.15: Intersection of elongated density distribution by the Ewald sphere at different places results in recording of diffraction spots which have the same Miller indices but different intensities ($I_1 \neq I_2$).

.

4.9 Conclusion

The data presented in this chapter show that collection of good quality data is a very challenging, labour intensive, and complicated task, both in terms of conducting the experiment and in data processing. The latter in particularly is non-standard and requires a lot of manual input and iterations in finding the tilt geometry of a crystal. Nevertheless, we were able to collect a 3D dataset from multilayered catalase crystals. In spite of the fact that completeness of the data is only 60-70 %, the data can be used for the next step of reconstruction – solving the phase problem.

Chapter 5

Molecular Replacement Theory

The ultimate aim of an X-ray or EM diffraction experiment is the computation of the electron density function $\rho(x, y, z)$ according to the Fourier synthesis given by equation:

$$\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l F_{hkl} e^{-2\pi i (hk + ky + lz)} \qquad (5.1)$$

It should be noted, that the structure factors F_{hkl} are complex numbers. To solve the equation 5.1, we need the values of all the structure factors F_{hkl} in both magnitude and phase. One of the fundamental problems in macromolecular crystallography is the *phase-problem*, which derives from the fact that only structure factor amplitudes can be measured during a diffraction experiment. This is unfortunate, since it can be shown that the phases are more important for the resulting features of the electron density map than the amplitudes (*66*). Various ingenious techniques have been invented to circumvent this tremendous problem, including direct methods, heavy atom methods, and molecular replacement. This chapter will introduce the fundamentals of the molecular replacement technique which was used to solve the phase problem in these studies.

5.1 Molecular Replacement

5.1.1 General Idea

Molecular replacement methods were first developed by Rossmann and Blow (*67*). It exploits the presence of non-crystallographic symmetry to obtain phase information and reduce phase uncertainties. The method was extended by Bricogne (*68*) for real-space symmetry averaging to improve the calculated electron density. Molecular replacement (MR) is based on the simple idea to find a molecule in different places in a unit cell by correlation techniques. More recently, the term "molecular replacement" is used for the process of finding a structurally homologous molecule in a different crystal form. If the structure of a molecule related to that under study is already known, it can be used to arrive at starting phases simply by positioning of the known molecule into the unit cell of the unknown.

The success of molecular replacement in macromolecular crystallography depends largely on the similarity between the search model and the protein of interest. It is assumed that there is an inverse correlation between the root-mean-square deviation of atomic positions and the percentage of sequence identity. This statement cannot be proven rigorously, but in practice this approach is often very useful for models with high sequence homology. Clearly, the higher the sequence identity the easier the MR. However, it is impossible to give an absolute percentage of sequence homology at which MR will no longer work. As a rule of thumb, if homology is about 80-90 %, molecular replacement solutions could be found easily. If homology is about 40-50%, it is still possible to solve molecular replacement, but optimisation of the parameters for rotation and translation search is required. If the model shares less than 40% sequence identity with the unknown structure, it becomes progressively more difficult to find the right position of the molecule in the unit cell. The largest problem with searches using low structurally homology probes is that once the solution is obtained, the phases will be poor estimates of the true phase, and there will be a high phase bias towards the probe model, making it difficult to refine the correct structure. This bias is the main drawback of the molecular replacement method.

In addition to similarity, the extent to which the search model represents the molecule of interest is also important. If the search model is very similar, but represents only 10% of the complete protein under study it will be harder to find a solution. Another problem that has come up is that sometimes there are a few copies of the molecule in the asymmetric unit. In this case, one solution would be to use MR for each individual copy.

5.1.2 Patterson Function

Traditional molecular replacement methods are based on the properties of the Patterson function, which is defined as

$$p(u,v,w) = \frac{1}{V} \sum_{h} \sum_{k} \sum_{l} \left| F_{hkl} \right|^2 e^{2\pi i(hu+kv+lw)} \qquad (5.2)$$

The Patterson function is simply the Fourier transform of the measured intensities. This may be obtained directly from an experiment, and so the Paterson function $p(u, v, w)$ may be computed immediately without phase information. A Patterson function of a trial atomic model can also be computed and compared to the observed Patterson function. When the model is oriented correctly and placed in the correct position in the unit cell, the two Pattersons should be similar. In this simplistic approach, the molecular replacement problem has six dimensions (three parameters to specify orientation and three to specify position), which results in a very large computational problem. Fortunately, the comparison of the two Pattersons can be divided into parts that are sensitive to only some of the six parameters.

The usual strategy is to look at these parts separately, reducing the dimensionality of the problem. One way to divide the problem up is to determine first the orientation (three parameters) by the *rotation function* and then the translation (also three parameters) by the *translation function*. It means that the first step in data processing is the search for an orthogonal transformation Ω, which allows to determine the relative orientation between the model and the unknown molecule. The second step includes the search of the

translation vector **v**, which determines the location of the molecule within the unit cell. Usually, the orthogonal transformation Ω: *(x, y, z)* $\rightarrow$ *(x', y', z')* is expressed by Euler angles or spherical angles.

There are several conventions for Euler angles, depending on the axes about which the rotations are carried out. Although the so-called "*x*-convention" is the most common definition, the CCP4 package uses the ZYZ convention (Figure 5.1). On this convention the rotation is described as three consecutive rotations:

1. Rotate around the Z axis by alpha degrees
2. Rotate around the new position of the Y axis for beta degrees
3. Rotate around the newest Z axis position for gamma degrees
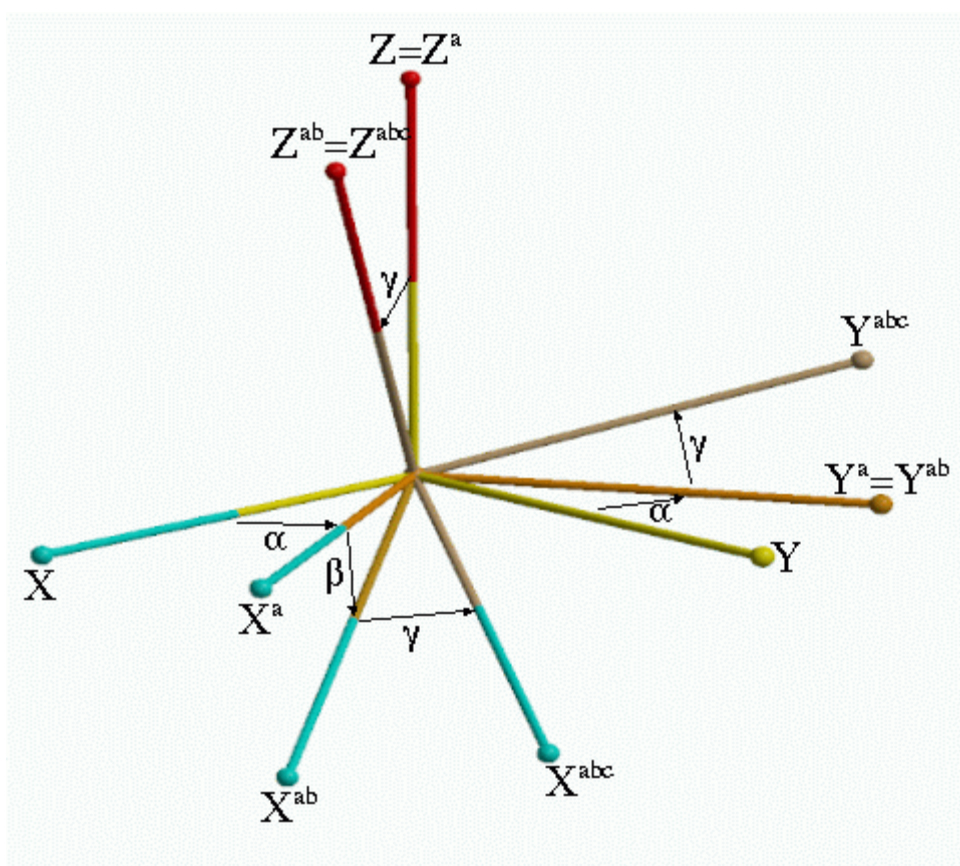
Figure 5.1: Definition of the Euler rotation angles used in the CCP4 program package.
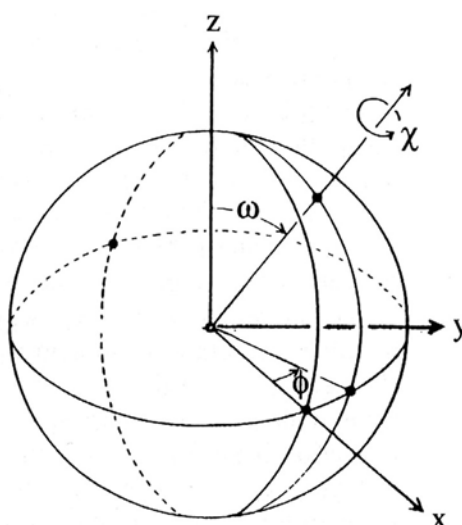
Figure 5.2: Definition of the polar rotation angles used in the CCP4 program package.

The spherical angles define the direction of a rotation axis and the magnitude of the angle to rotate. In this definition, ϕ and ω define the direction of the rotation axis, and χ defines the rotation angle (see Figure 5.2). This definition is especially useful for the *self-rotation function* where distinct rotation angles are expected; e.g. 180 degrees for a dimer, 120 degrees for a trimer (Chapter 5.5).

5.1.3 Construction of a Patterson Function

 The construction of a Patterson function can be regarded as consecutively placing each atom at the origin and drawing vectors from such an atom to all other atoms in the unit cell. For a structure with N atoms there are N(N-1) non-origin peaks. The value assigned to the vector is proportional to the product of the atomic numbers at the end of the vector. Thus, the height of the origin peak in a Patterson map is proportional to the sum of the squared atomic numbers (due to the null-vectors from each atom to itself).

A Patterson map contains information of two types of vectors:

1. *Intramolecular* or self-Patterson vectors of pairs of atoms in the same molecule. These vectors are relatively short and are thus clustered around the origin. Intramolecular vectors depend only on the orientation of the molecule, and not on its position in the cell, so these can be exploited in the rotation function (Chapter 5.2).

2. *Intermolecular* or cross-Patterson vectors, which are generally longer than self-vectors. The cross-Patterson vectors depend both on the orientation of the molecule and on its position. So, once the orientation is known, these can be exploited in the translation function (Chapter 5.3).

5.2 Rotation Function

5.2.1 The Rossmann-Blow Rotation Function

Suppose that the Patterson function is now superimposed onto a rotated version of itself. There will be no particular agreement except when one set of self-Patterson vectors of one molecule has the same orientation as the self-Patterson vectors from the other molecule. In this position the agreement or "overlap" between the two will be at a maximum. The function describing this overlap has been termed by Rossmann and Blow as the *Rotation function* (*67*), and is defined as follows:

$$R(\Omega) = \int_U P_x(\mathbf{r})P_m(\Omega\mathbf{r})d\mathbf{r} \qquad (5.3)$$

where $\mathbf{P}_x(\mathbf{r})$ is the observed Patterson function, $\mathbf{P}_m(\Omega\mathbf{r})$ is the calculated Patterson function dependent on Eulerian angles α, β, γ, $\mathbf{r}$ is a Patterson vector, U is the appropriate volume of integration, usually spherical, centred at the origin. A maximum in the rotation function $R(\Omega)$ indicates a potentially correct orientation of the search probe in the target cell. The rotation function can be rewritten as:
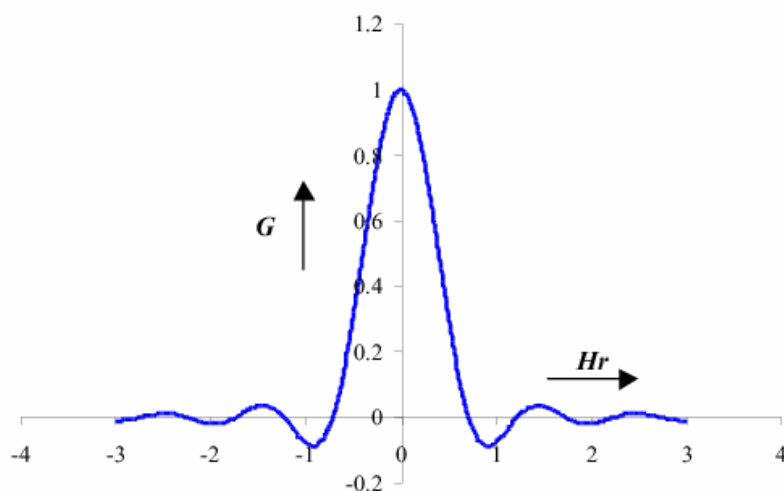
Figure 5.3: The Interference function G.

$$R(\Omega) = \frac{U}{V^3} \sum_h \sum_{h'} \left| \mathbf{F}_x(\mathbf{h}) \right|^2 \left| \mathbf{F}_m(\mathbf{h'}) \right|^2 G_{hh'} \qquad (5.4)$$

where $\mathbf{h}=(h, k, l)$ are Miller indexes, $\mathbf{F}_x(\mathbf{h})$ and $\mathbf{F}_m(\mathbf{h})$ are the structure factors of the target and the probe molecules respectively, $G_{hh'}$ is the so-called *interference function*. For an integration sphere of radius r it equals:

$$G_{hh'} = \frac{3(\sin 2\pi Hr - 2\pi Hr \cos 2\pi r)}{(2\pi Hr)^3} \qquad (5.5)$$

where $H = \left| \mathbf{h+h'} \right|$. The maximum value of the interference function is 1.00 (see figure 5.3), and it is never greater than 0.086 outside the range $-0.725 < x < 0.725$. Hence all terms in equation 5.5 for which $\left| Hr \right| > 0.725$ may be neglected if U is assumed to be a sphere with radius r.

Two major problems affect the efficiency of the Rossmann-Blow rotation function:

- the relatively long CPU time required to compute it;
- the lack of precision resulting from truncation errors associated with the G_{hh} function.

5.2.2 The Crowther Fast Rotation Function

Originally, the rotation function was computed by directly comparing two Pattersons or lists of peaks in those Pattersons. Such rotation functions are still sometimes used, but most of the time the fast rotation function developed by Crowther is used (*69*). He realised that the rotation function could be computed quickly with Fourier transforms, if the Patterson functions were expressed in terms of spherical harmonics. The Crowther fast rotation function makes use of spherical coordinates instead of Cartesian ones. The Patterson function is expanded in terms of the spherical harmonics:

$$S_{lmn}(r,\theta,\phi) = jl(k_{ln}r)Y_l^m(\theta,\phi) \qquad (5.6)$$

and

$$P_A(r,\theta,\phi) = \sum_{lmn} a_{lmn} S_{lmn} \qquad (5.7)$$

$$P_B(r,\theta,\phi) = \sum_{l'm'n'} b_{l'm'n'}^* S_{l'm'n'}^* \qquad (5.8)$$

where *jl* is the normalized spherical Bessel function of order *l* and k_{ln} is chosen such that $jl(k_{ln}R)=0$. Y_l^m is the normalized spherical harmonics function. The expansion coefficient a_{lmn} can be calculated from

$$a_{lmn} = \sum_h F_h^2 T_{lmn}(h) \qquad (5.9)$$

where T_{lmn} is the Fourier transform of S_{lmn}.

With such an expansion for both Patterson functions and utilizing the special properties of the spherical harmonics, the rotation function can be rewritten as:

$$R(\theta_1,\theta_2,\theta_3) = \sum_m \sum_{m'} f_{mm'}(\theta_2)e^{-im\theta_1}e^{-im'\theta_3} \qquad (5.10)$$

where the rotation is represented by a set of Eulerian angles (θ_1, θ_2, θ_3), and the function $f_{mm'}$ is only dependent on θ_2. Eq. (5.10) is a two-dimensional Fourier transform over θ_1 and θ_3. Therefore, for any given θ_2, the equation can be evaluated by the fast Fourier transform (FFT) technique, which greatly accelerating the rotation function calculation.

In comparison to the Rossmann-Blow rotation function, the Crowther fast rotation function has the following advantages:

- It is about 100 times faster in terms of CPU usage;
- The calculation is more precise, since there is no truncation of the G function for large $|F|^2$;
- It is easier to take the symmetry of the Patterson function into account. For example, if we have a symmetry axis of the N-th order along the z axis, then the non-zero coefficients among the coefficients a_{lmn} will be only $a_{lkN\,n}$ (k=0, 1, 2...);
- The origin peak can be removed in a very simple way by omitting the a_{000} coefficient from the calculation.

5.2.3 Rotation Function in *AMoRe*

The rotation functions used in the studies of this thesis were calculated with the program package *AMoRe* (*70*). The rotation function in *AMoRe (71)* is based on the fast rotation function, using spherical harmonics and Bessel function expansions. The adaptations made by Navaza permit more accurate computation of the rotation matrices (*72*), and enhance the resolution of the rotation function peaks. Removal of the origin (usually performed by using an inner limit for the integration radius) is achieved by excluding the lower order spherically symmetrical Bessel functions.

AMoRe uses an uncommon philosophy at the stage of the structure factor calculation from the model electron density. The model is rotated so that its principle axes of inertia are aligned with the cell axes and then placed, with its centre of mass, at the origin of a P1 cell with linear dimensions around four times the size of the model. Continuous Fourier

coefficients are then calculated on a fine grid. Therefore the problem of the need for sampling at non-integral reciprocal lattice points during the calculation of the rotated Patterson function, does not occur. The structure factors can be interpolated from the continuous set, rather than being calculated for every rotation. This procedure results in a small error: only 1% in terms of crystallographic R-factor for the situation of a quadruple-size box. For a double-size box this discrepancy would be 8% (*73*).

The program outputs a list of rotation function peaks with Eulerian angles α, β, γ and a correlation coefficient (CC). Because the model has been placed at the origin and rotated, the output is not in straightforward Euler angles. It needs to be shifted back, away from the origin.

5.3 Translation Function

Once a rotation solution has been found, it can be tried to find a solution of the translation function. This step is not as robust as the rotational search. It is possible to obtain a correct rotation solution and still not find a translation solution. The main concept of the translation function is the same as for the rotation function. The correlation between the observed intensities and the Patterson cross-vectors of the symmetry-related molecules of the probe are calculated as it is moved within the cell. When the probe is correctly positioned, the translation function should have peaks at values corresponding to the translation vectors between the symmetry related molecules. There are several forms of the translation function, which will be discussed in the following sections.

5.3.1 The T and T1 Translation Functions

Translation functions are calculated by moving properly oriented probes around the unit cell volume. Crowther and Blow (*74*) have initially represented the T translation function as:

$$T_{jk}(\mathbf{v}) = \int_V P_{jk}(\mathbf{u}, \mathbf{v}) \, P_{obs}(\mathbf{u}) d\mathbf{v} \qquad (5.11)$$

$P_{jk}(\mathbf{u},\mathbf{v})$ is the cross-Patterson function of the model structure in which two molecules (j and k) are related by crystallographic symmetry. $\mathbf{v}$ is the intermolecular vector between the local origins of these two molecules. $P_{obs}(\mathbf{u})$ is the Patterson function of the unknown structure. This function effectively searches one pair of asymmetric units, *i.e.* one crystallographic symmetry element, at a time. Since the self vectors confuse the picture, they can easily be removed by subtraction, resulting in the T1 function:

$$T_{jk}1(\mathbf{v}) = \int_V \left[P_{obs}(\mathbf{u}) - \sum_i P_{ii}(\mathbf{u}) \right] P_{jk}(\mathbf{u}, \mathbf{v}) d\mathbf{u} \qquad (5.12)$$

5.3.2 The TO Translation Function

The solution of the translation function T1 only represents one symmetry operator; all other symmetry information (including non-crystallographic symmetry) is lost. An improved translation function has been derived, resulting in peaks for all possible intermolecular vectors. The model now has as many molecules in the cell as the unknown structure. It is more convenient to express this translation function in terms of the translation vector $\mathbf{t}$ (a vector between the model and the unknown molecule). The expression contains the calculated Patterson for the correctly oriented model with unknown translation:

$$TO(\mathbf{t}) = \int_V P_{obs}(\mathbf{u}) \, P_{calc}(\mathbf{u}, \mathbf{t}) d\mathbf{u} \qquad (5.13)$$

5.3.3 The T2 Translation Function

Both improvements from the T1 and T0 translation function can be incorporated in the most commonly used T2 translation function:

$$
T2\ (\mathbf{t}) = \int_V \left[P_{obs}(\mathbf{u}) - \sum_i P_{ii}(\mathbf{u}) \right] \left[P_{calc}(\mathbf{u},\mathbf{t}) - \sum_i P_{ii}(\mathbf{u}) \right] d\mathbf{u} \qquad (5.4)
$$

The T2 translation function applies phase shifts so that all peaks coincide at the same place, defined by a single translation vector **t** for the reference molecule. It results in a much greater signal-to-noise ratio than for a single T1 function.

Translation function calculations can be improved by:

- Including existing experimental phase information from isomorphous replacement studies, calculating the so-called "phased translation function" (*75*);
- Including information of a partial model by either adding or subtracting partial Patterson functions appropriately during rotation or translation function calculations (*76*);
- Improving the search model and its orientation before the translation search. Brünger (*77*) has suggested the procedure of Patterson Correlation refinement, minimising a target function based on the correlation between observed and calculated Patterson functions and including an empirical energy term to restrain the geometry of the model. The model can be divided into domains treated as rigid bodies, while the domains are allowed to move with respect to each other;
- fixing one molecule in the asymmetric unit while searching for non-crystallographically related molecules (as in *AMoRe*).

5.3.4 Translation Search

An alternative to the translation function is a R-factor search where the molecule is moved on a grid and at each point the R-factor between the calculated probe amplitudes and the observed unknown structure amplitudes is calculated (eq. 5.15).

$$R = \frac{\sum\limits_{hkl} \left\| F_{obs} \right| - k \left| F_{calc} \right\|}{\sum\limits_{hkl} \left| F_{obs} \right|} \qquad (5.15)$$

During these R-factor search calculations, the Fourier transform of the molecule only needs to be calculated once. All translations of the probe molecule correspond to a pure phase shift which can be applied directly.

A more robust method is to calculate the correlation function since it is insensitive to structure factor scaling errors which usually are present in electron diffraction experiment data. The translation search takes more computer time than the translation function, but it is less prone to error and it will be the method of choice as computers become faster. Several proteins that could not be solved using the translation function were solved using the correlation search. In this method the standard linear correlation coefficients (SLCC) are calculated according to equation 5.16:

$$SLCC = \frac{\sum\limits_{hkl} \left(\left| F_{obs} \right|^2 - \left\langle \left| F_{obs} \right|^2 \right\rangle \right) \left(\left| F_{calc} \right|^2 - \left\langle \left| F_{calc} \right|^2 \right\rangle \right)}{\left[\sum\limits_{hkl} \left(\left| F_{obs} \right|^2 - \left\langle \left| F_{obs} \right|^2 \right\rangle \right)^2 \sum\limits_{hkl} \left(\left| F_{calc} \right|^2 - \left\langle \left| F_{calc} \right|^2 \right\rangle \right)^2 \right]^{1/2}} \qquad (5.16)$$

5.3.5 Translation Function in *AMoRe*

AMoRe uses conventional T0 translation functions (*78*), with the option to incorporate information about previously located models (*70*), which enables the positioning of

additional molecules in an asymmetric unit. The program outputs the α, β, γ angles, a translation vector (x, y, z) , a R-factor and a correlation coefficient.

5.4 Factors Affecting the Molecular Replacement Solution

Solving a structure by the molecular replacement method is not always a straightforward task. Depending on the complexity of the problem, a number of important factors must be taken into account in order to ensure a correct structure determination. In this section the description of the factors affecting the molecular replacement solution is shortly discussed.

5.4.1 Rotation Function Integration Radius

Usually, the volume of integration U (eq. 5.3) is a spherical shell which is described by the minimal radius r_{min} and the maximal radius r_{max}. In the integration over the spherical shell, a region near the origin is typically omitted to exclude the large origin peak of the Patterson, which would otherwise add a large constant term. To exclude the origin peak, the radius r_{min} should be larger than 2 Å. The choice of the maximal radius r_{max} depends on the dimensions of the molecule and packing of the unit cell and needs to be large enough to exclude cross vectors. None of the intramolecular vectors is longer than the longest dimension of the protein so the maximal radius of integration should never be larger than this. The rule of thumb is to use the maximal radius of integration that is approximately 75% of the largest dimension. This will cover most of the intramolecular vectors without too much contamination by intermolecular vectors.

5.4.2 Resolution

MR is sensitive to both the low and high resolution limits which are used during the analysis. Higher resolution reflections (above about 3.5 Å) may differ markedly even between homologous structures as they reflect the precise conformations of residues. Lower resolution reflections reflect the grosser features of the structure, such as the relative positions of secondary structure elements. Very low resolution reflections, below about 10 Å, are heavily influenced by the crystal packing and the arrangement of solvent and protein. This gives systematic errors that are stronger at low resolution. In X-ray practice, the very low resolution data is not normally useful. Thus, the resolution range used for rotation searches is usually within 10-3.5 Å, with 8-4Å being common. However, several ranges should be tried; the usable range for electron crystallography will be discussed in Chapter 6.

5.4.3 Number of Molecules in the Asymmetric Unit

The crystal form most favourable for the molecular replacement method is the one with only a single molecule in the asymmetric unit. The presence of multiple copies of the protein in the asymmetric unit reduces the signal-to-noise ratio of the correct peaks of the rotation and translation functions. However, there are many cases in which the structure was solved by molecular replacement method where there are several molecules in the asymmetric unit.

5.4.4 Data Completeness

It is very important that the experimental data are as complete as possible, ideally 100%. Also, the data should be of high quality. Possible problems include systematically missing regions (in the case of an EM diffraction experiment it is a missing cone effect (see Chapter 4)), ice ring background, detector overloads, etc.

5.4.5 The Use of Normalised Structure Factors *E(h)*

The normalised structure factors are given by the following formula:

$$E(\mathbf{h}) = \frac{\mathbf{F}\,(\mathbf{h})}{\sqrt{\left|\left\langle \mathbf{F}\,(\mathbf{h}) \right\rangle\right|^2}} \qquad (5.17)$$

where F(**h**) are structure factor amplitudes. The average value of $\mathbf{E}^2$ is 1. The physical meaning of the normalised structure factors is that they are the structure factors corresponding to a structure of stationary point atoms. The use of normalised structure factors in the molecular replacement calculations can make the correct solutions more obvious and is one of the ways of sharpening the Patterson peaks.

5.5 Non-crystallographic Symmetry and the Self-rotation Function

To solve a molecular replacement problem, knowledge about the internal symmetry, or non-crystallographic symmetry (NCS), of the object can be very helpful. If there are N molecules in the asymmetric unit, it means that a search model consisting of a monomer represents only 1/N of the total atoms in the asymmetric unit. Therefore it will be harder to find the rotation function solution. However, many macromolecular crystals contain assemblies of the macromolecules obeying a simple (local, or non-crystallographic symmetry) point group symmetry. Non-crystallographic symmetry is a symmetry that exists locally within the asymmetric unit of the crystal. For example, the protein can be crystallized as a dimer in the asymmetric unit so that there is a dyad axis relating the halves of the dimer that is not coincident with a crystallographic 2-fold. If there is information about NCS, a special version of the rotation function, the so-called *locked rotation function*, cab be used to determine the orientations of all symmetry related elements of the

point group at the same time. Furthermore, NCS can be used to produce averaged maps in which noise tends to be smaller.

In order to discover non-crystallographic symmetry a *self-rotation function* should be calculated. In the self-rotation function, the Patterson of the unknown structure is rotated with respect to itself. Maximum overlap of these Patterson maps will occur at zero rotation and at rotations representing the non-crystallographic symmetry. It is very useful to calculate a self-rotation function in spherical angles. By looking at the section of the outermost nested angle (usually χ, Fig. 5.2), the position of these non-crystallographic symmetry elements can be found. For example, section $\chi=180^\circ$ contains information about 2-folds. It means that when the second copy of the molecule is rotated by 180 degrees with respect to the first, it will be superimposed in the direction of the 2-fold and give a peak in the self-rotation function. In general, if NCS is a *n*-fold rotation axis, sections $360/n$ must have well-defined maxima. For example, peaks on the section $\chi=120^\circ$ are due to 3-folds; $\chi=90^\circ$ are due to 4-folds; $\chi=72^\circ$, 5-folds and $\chi=60^\circ$, 6-folds.

Chapter 6

Molecular Replacement Results

This chapter describes each step for finding the correct molecular replacement solution using EC-diffraction data from multilayered crystals of bovine catalase. We solved the rotation and translation functions and found the orientation and position of a catalase tetramer in the unit cell. Since the experimental setups do not allow to collect a complete dataset, the effects of data completeness on a self-rotation function and refinement of the model are analyzed and discussed.

6.1 Effect of Incomplete Data on Molecular Replacement Solutions

As discussed in the previous chapter, data completeness is one of the possible parameters that can affect the molecular replacement solution. Thus, the data collection strategy should ensure and optimize a way of collecting a dataset with a very high degree of completeness. Even the loss of a few percent of the unique data can have a highly detrimental effect on the molecular replacement solutions, density modification procedures, refinement and electron density map interpretation. Usually, in monochromatic X-ray studies datasets with a completeness of at least 80-90 % have been used for molecular replacement.

In the case of multilayered crystals diffraction data must be recorded from differently tilted crystals. Since the maximal tilt angle for modern cryoholders is limited to ±60º, the

maximal completeness which could be achieved in our studies is only 87% (see Chapter 4). The systematic data loss is known as the *missing cone effect*. However, the collection of the 87% complete dataset from bovine catalase multilayered crystals is extremely time consuming. In our studies we were able to collect a dataset with a completeness of only 50-70 % for different resolution shells. So, the first principle question that we had to answer was: Is it possible to find a correct molecular replacement solution for such incomplete data?

Bovine liver catalase has been crystallized in different crystal forms. Ko *et al.* (*58*) have used crystals of the orthorhombic space group $P2_12_12_1$ with a=87.8 Å, b=140.6 Å, and c=232.4 Å for X-ray structure determination. The structure was solved and the coordinates and experimental structure factors amplitudes were deposited in the Protein Data Bank (PDB) under the access code 4BLC. The multilayered catalase crystals used in our work also belong to the orthorhombic space group $P2_12_12_1$, but with different unit cell parameters a=69.0 Å, b=173.0 Å, and c=206.0 Å. Although the crystal packing for these two crystal forms is different (as evident by the different unit cell parameters), the behaviour of the rotation and translation functions should be similar[14]. We used structure factors amplitudes from 4BLC as a reference dataset to test our programs and find optimal parameters for molecular replacement.

To test the effect of the missing cone in our data, we used structure factor amplitudes from 4BLC to investigate the effect of incomplete data on molecular replacement solutions. The results are summarized in Tables 6.1-6.6. Six datasets with different completeness were created:

1) "Xray_100". This dataset contained all experimentally collected reflections.
2) "Xray_90". 10% of randomly distributed reflections were removed from Xray_100.
3) "Xray_80". 20% of randomly distributed reflections were removed from Xray_100.
4) "Xray_70". 30% of randomly distributed reflections were removed from Xray_100.
5) "Xray_60". 40% of randomly distributed reflections were removed from Xray_100.

---

[14] Both the parameters (Patterson radius, probe unit cell, resolution cut-off) and the behaviour of the signal/noise ratio for rotation function and a correlation coefficient for the translation function should be similar.

6) "Xray_EMcompl". This dataset contained reflections from Xray_100 which had the same Miller indexes as reflections from electron diffraction dataset (see Chapter 4).

Table 6.1: List of the first 6 peaks of the MR solutions obtained with a tetramer of bovine liver catalase (4BLC) and a Xray_100 dataset (the two correct solutions are highligthed).

Peak Number	α	β	γ	T_x	T_y	T_z	Corr. Coef.	R-factor
1	**29.94**	**86.41**	**189.49**	**0.2595**	**0.3084**	**0.3169**	**78.8**	**27.0**
2	**29.83**	**86.50**	**9.46**	**0.2595**	**0.3084**	**0.3169**	**75.6**	**29.2**
3	91.40	15.18	329.55	0.4342	0.1782	0.1517	6.3	53.0
4	92.60	164.69	210.44	0.4500	0.2223	0.1247	5.7	53.5
5	167.40	68.35	116.30	0.3774	0.1703	0.0875	5.7	53.6
6	17.93	40.22	229.91	0.0810	0.2810	0.1316	5.6	53.5

Table 6.2: List of the first 6 peaks of the MR solutions obtained with a tetramer of bovine liver catalase (4BLC) a Xray_90 dataset (the two correct solutions are highligthed).

Peak Number	α	β	γ	T_x	T_y	T_z	Corr. Coef.	R-factor
1	**29.93**	**86.32**	**189.53**	**0.2595**	**0.3083**	**0.3169**	**78.6**	**27.2**
2	**29.81**	**86.44**	**9.47**	**0.2594**	**0.3083**	**0.3170**	**75.9**	**29.0**
3	92.53	31.44	218.90	0.4535	0.2878	0.3629	7.2	53.1
4	84.51	13.76	169.58	0.0978	0.3213	0.0805	6.9	53.3
5	46.89	45.18	46.19	0.2103	0.3398	0.4845	6.8	53.5
6	47.54	132.85	130.99	0.1995	0.2258	0.4772	6.4	53.7

Table 6.3: List of the first 6 peaks of the MR solutions obtained with a tetramer of bovine liver catalase (4BLC) and a Xray_80 dataset (the two correct solutions are highligthed).

Peak Number	α	β	γ	T_x	T_y	T_z	Corr. Coef.	R-factor
1	**29.93**	**86.31**	**189.53**	**0.2595**	**0.3083**	**0.3169**	**78.8**	**26.9**
2	**29.83**	**86.42**	**9.46**	**0.2595**	**0.3083**	**0.3170**	**75.6**	**29.1**
3	15.15	110.87	116.34	0.1710	0.0795	0.2993	6.8	53.6
4	45.24	133.97	131.50	0.1935	0.2378	0.4697	6.5	53.7
5	44.59	46.19	48.48	0.2033	0.3231	0.4573	6.2	54.0
6	11.40	67.49	241.98	0.2868	0.4755	0.0916	6.1	53.6

Table 6.4: List of the first 6 peaks of the MR solutions obtained with a tetramer of bovine liver catalase (4BLC) and a Xray_70 dataset (the two correct solutions are highligthed).

Peak Number	α	β	γ	T_x	T_y	T_z	Corr. Coef.	R-factor
1	**29.89**	**86.36**	**189.58**	**0.2595**	**0.3083**	**0.3170**	**78.3**	**27.2**
2	**29.79**	**86.48**	**9.49**	**0.2595**	**0.3083**	**0.3171**	**75.4**	**29.3**
3	85.60	165.80	188.60	0.1695	0.1249	0.0037	7.0	52.6
4	87.40	13.75	168.60	0.0997	0.3013	0.0805	6.8	53.6
5	14.02	68.66	61.49	0.3662	0.2963	0.1672	6.7	53.6
6	92.60	14.88	148.40	0.3318	0.2173	0.2991	6.2	53.6

Table 6.5: List of the first 6 peaks of the MR solutions obtained with a tetramer of bovine liver catalase (4BLC) and a Xray_60 dataset (the two correct solutions are highligthed).

Peak Number	α	β	γ	T_x	T_y	T_z	Corr. Coef.	R-factor
1	**29.88**	**86.19**	**189.49**	**0.2596**	**0.3083**	**0.3172**	**77.6**	**27.8**
2	**29.79**	**86.34**	**9.36**	**0.2596**	**0.3083**	**0.3173**	**75.2**	**29.5**
3	135.01	46.14	130.75	0.0796	0.3583	0.1571	6.9	52.4
4	84.22	14.34	351.24	0.0197	0.1584	0.2937	6.7	53.5
5	93.26	14.03	189.63	0.0202	0.3893	0.2485	6.6	53.6
6	133.05	145.11	28.41	0.4191	0.2223	0.2442	6.6	53.8

Table 6.6: List of the first 6 peaks of the MR solutions obtained with a tetramer of bovine liver catalase (4BLC) and a Xray_EMcompl (the two correct solutions are highligthed).

Peak Number	α	β	γ	T_x	T_y	T_z	Corr. Coef.	R-factor
1	**29.94**	**85.89**	**189.11**	**0.2595**	**0.3083**	**0.3175**	**74.7**	**29.3**
2	**29.90**	**86.07**	**9.00**	**0.2595**	**0.3083**	**0.3177**	**71.7**	**31.2**
3	91.53	34.55	216.57	0.4324	0.3040	0.3508	7.2	53.3
4	44.52	46.10	228.54	0.3079	0.2965	0.0095	6.8	53.6
5	130.03	34.85	321.05	0.0808	0.2772	0.0399	6.8	53.7
6	114.69	83.39	176.08	0.0713	0.1435	0.4837	6.7	53.1

The translation function which was calculated using the catalase tetramer (two dimers) shows two highest peaks for all datasets independent of the varying completeness. These two orientations are related by 180° as described in Chapter 6.5 and are symmetry equivalent solutions reflecting the NCS of the tetramer. The correct solutions have correlation coefficients higher than 70 % and a low R-factor. In addition, there is a big gap between the correct solutions and background noise.

These results clearly indicate that the correct solution of the molecular replacement analysis could be found even with low completeness data. This means, that incompleteness itself is not an insurmountable obstacle to find molecular replacement solutions and, therefore, we expect to solve the phase problem using our EM dataset.

6.2 Model

The choice of the correct model is very important for the molecular replacement method. It is known that bovine liver catalase consists of a tetramer and that its molecular weight is 232,000 (*79*). It has been reported that a diluted solution of catalase is unstable (*80*) because the tetramer may easily dissociate into two dimers. To estimate the number of molecules in the asymmetric unit, we calculated the Matthews number V_M (*81*):

$$V_M = \frac{V_{cell}(\text{Å}^3)}{M n_{asymun} n_{mol}} \qquad (6.1)$$

where *M* is the protein molecular weight in Daltons; V_{cell} is the unit cell volume; n_{asymun} is the number of asymmetric units; n_{mol} is the number of molecules in the asymmetric unit (asu)[15].

The results of calculation of V_m for the two crystalline forms with different numbers of catalase monomers in the asymmetric unit are summarized in Table 6.7.

Table 6.7: The Matthews number.

Number of catalase monomers in asu		2	4	6	8
Crystal form I (X-ray)	V_M	6.5	**3.2**	2.2	1.6
	Solvent fraction (%)	81	**62**	43	24
Crystal form II (EM)	V_M	5.6	**2.8**	1.9	1.4
	Solvent fraction (%)	77	**56**	32	12

---

[15] An asymmetric unit is defined as the smallest unit from which the crystal structure can be generated by making use of the symmetry operations of the space group.

Usually, the value of V_m ranges from 1.7 to 3.5 $Å^3$/Da depending on the solvent fraction of the crystal. However, Matthews coefficients of about 2.2 $Å^3$/Da are more typical for protein crystals with 50% solvent content. The x-ray structure of catalase (4BLC, crystal form I) indicates that there are four catalase monomers in the asu which form a tetramer. For crystal form II, the presence of four monomers in the asu shows the most optimal values of V_m and the solvent fraction. We therefore assumed that the monomers are arranged in a similar manner as in crystal form I and also form a tetramer. So, the tetramer molecule (4BLC) was chosen as a search probe for molecular replacement.

6.3 Using the 'Automated Molecular Replacement' (*AMoRe*) Package for Solving the Structure of Catalase by EC

We choose the Automated Molecular Replacement software package (*AMoRe*) (*70*) for solving the molecular replacement problem. The flow-chart of the *AMoRe* procedure is shown in figure 6.1. This package can use both structure factor amplitudes and normalised structure factor amplitudes in its calculations. The following sections describe the various steps and operations of the *AMoRe* package that were applied to the catalase MR problem using structure factor amplitudes from electron diffraction from multilayered crystals in the calculation.
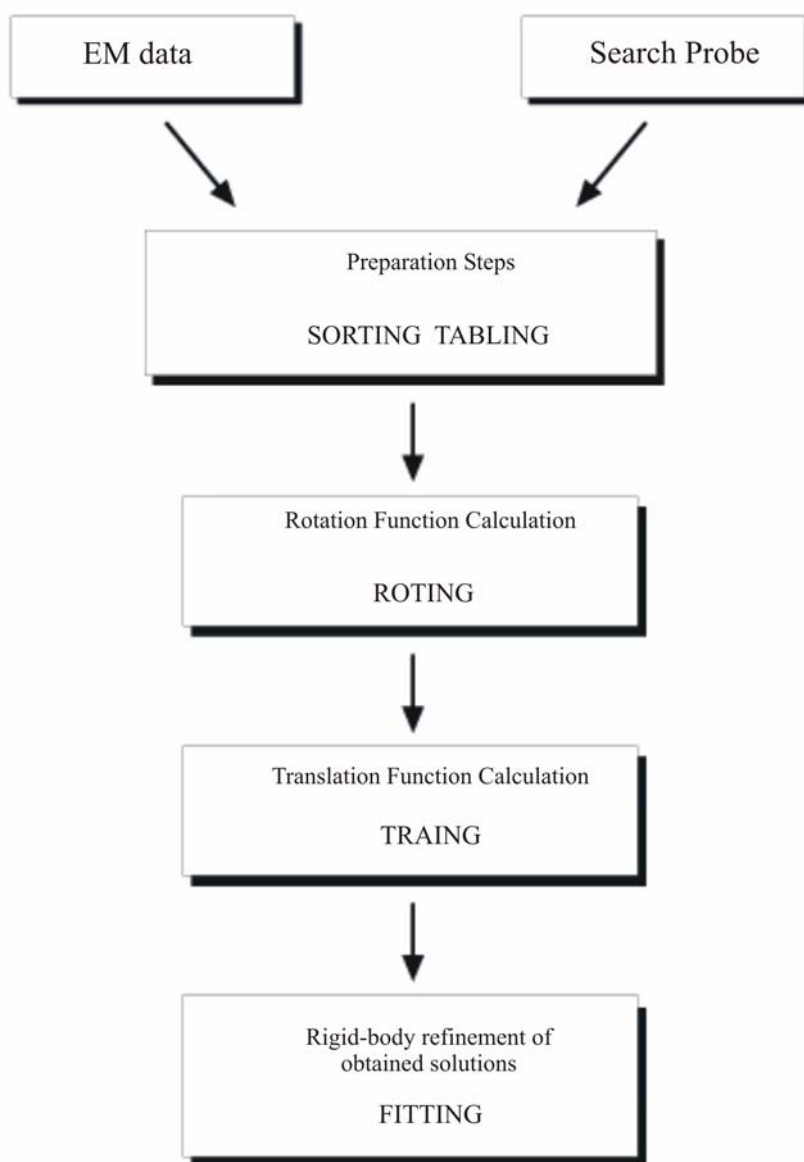
Figure 6.1: The *AMoRe* molecular replacement procedure flow-chart. The step-by-step guide of *AMoRe* practice. The individual programs used at each step are indicated in capital letters.

6.3.1 The Preparation Steps

Two preliminary programs must be run in order to format the data into a suitable representation. The first is SORTING, which transforms the measured intensities into the *AMoRe* internal format. The second program is TABLING, which calculates the continuous Fourier coefficients, corresponding to the search probes. The continuous

Fourier coefficients have been used in the structure factor calculation by interpolation (*70*). These coefficients are obtained by Fourier transforming the electron density based on the atomic coordinates of each search probe, which is placed within a cell with linear dimensions about four times the size of the molecule. To avoid creating an unnecessary large cell, caused by an unfavourable positioning of the probe, the molecule is shifted and rotated so that its centre of mass is at the origin and its principal axes of inertia are aligned with the cell axis. The resulting position is taken as the initial reference, and modified atomic coordinates of each probe are written out.

The default version of *AMoRe* is adopted for X-ray studies and the continuous Fourier coefficients are calculated using atomic scattering factor coefficients for X-rays. For molecular replacement from electron diffraction data atomic scattering factor coefficients for electrons must be used. In our studies we have used gaussian coefficients for electron scattering factors evaluated for 120 kV electrons by Ceska (*82*).

6.3.2 The Rotation Function Search

The next step is the calculation of the rotation function using the program ROTING. First, the structure factor amplitudes of a search model have to be computed in a P1 cell. This cell is substantially smaller than that used to compute the continuous Fourier coefficients. The size of this cell can be calculated in two ways proposed by Dodson and Navaza (*70*). The first way suggested by Dodson (*83*) is to define a cubic cell with the edges of double the maximum distance from the centre of mass, plus the integration radius, plus a small safety term. Navaza (*70*) advises to take the size of this cell equal to the smallest box containing the search probe, plus the integration radius, plus the resolution used for calculation. In any case the cell should be big enough to accommodate the molecule.

Next, the spherical harmonic coefficients for the target and probe Patterson distributions are computed. Finally, the rotation function is calculated. The result of the rotation function search, relative to the chosen molecular model, consists of a list of possible solutions sorted by the peak heights. Each output line contains the three Eulerian angles

which identify the different orientations of the probe molecule in the cell and the peak height.

6.3.2.1 The Integration Radius of the Search Probe

The choice of the integration radius in the calculation of rotation functions depends on the size of the search probe. As previously mentioned (Chapter 5), a value corresponding to 75% of the minimal diameter of the molecule normally represents a good approximation. The TABLING program outputs the size of the minimal box containing the search probe and the distance of the furthest atom from the centre of mass. These values, as well as the length corresponding to 75% of the minimal diameter for the catalase tetramer search probe, are listed in table 6.8.

Table 6.8: The size of the catalase tetramer molecule.

Probe	Minimal Box, Å	Maximal distance from centre of mass, Å	75% of minimal diameter, Å
4BLC	102.4 98.4 72.7	57.2	54.5

6.3.2.2 Results of the Rotation Search

The rotation function search clearly shows two peaks. The parameters for the calculation were: integration radius of 30-60 Å, different resolution limits from 20 to 4 Å, a cubic search probe cell with axis equal to 150 Å. No structure factor sharpening was used. In table 6.9 the first 10 peaks from the rotation function calculation are listed. Tables 6.10 and 6.11 show the effect of different resolution limits and radius of integration on the contrast between the correct solutions and background noise.

Table 6.9: The list of the first 10 peaks rotation function (the two correct solutions are highlighted).

Peak Number	α	β	γ	Peak height, σ
1	**20.57**	**105.04**	**188.04**	**11.6**
2	**20.70**	**105.12**	**8.04**	**11.4**
3	17.91	107.36	359.03	3.8
4	15.51	126.57	145.64	3.6
5	18.57	107.44	179.72	3.6
6	88.96	79.36	270.17	3.5
7	97.60	165.01	277.60	3.4
8	90.56	101.29	269.84	3.4
9	162.23	107.26	185.87	3.3
10	16.29	126.04	326.60	3.3

Resolution limits: 10 - 4 Å; probe unit cell: 150, 150, 150 Å; integration radius: 55 Å.

Table 6.10: The effect of the resolution limits on the signal of the correct solutions (the best combination of resolution limits is highlighted).

Resolution Limits, Å	Solution #1, σ	Solution #2, σ	The highest background peak, σ	Solution #1/noise	Solution #2/noise
20-4	16.6	16.5	8.9	1.9	1.9
18-4	16.2	16.1	9.1	1.8	1.8
16-4	15.2	15.0	8.6	1.8	1.7
14-4	15.3	15.3	7.3	2.1	2.1
12-4	12.9	12.6	6.4	2.0	2.0
10-4	**12.7**	**12.2**	**5.7**	**2.2**	**2.1**
8-4	11.2	11.1	5.9	1.9	1.9

The integration radius of 40 Å and the probe unit cell of 150, 150, 150 Å were used for calculations.

Table 6.11: The effect of the integration radius changes on the signal of the correct solutions (the best radius is highlighted).

Integration Radius, Å	Solution #1, σ	Solution #2, σ	The highest background peak, σ	Solution #1/noise	Solution #2/noise
25	12.2	11.5	9.6	1.3	1.2
30	13.4	12.7	8.3	1.6	1.5
35	13.0	12.2	6.6	2.0	1.8
40	12.7	12.2	5.7	2.2	2.1
45	12.5	12.4	4.7	2.7	2.6
50	12.4	12.3	4.2	3.0	2.9
55	11.6	11.4	3.8	3.1	3.0
60	**10.7**	**10.4**	**3.2**	**3.3**	**3.3**

Resolution limits: 10 - 4 Å; probe unit cell: 150, 150, 150 Å.

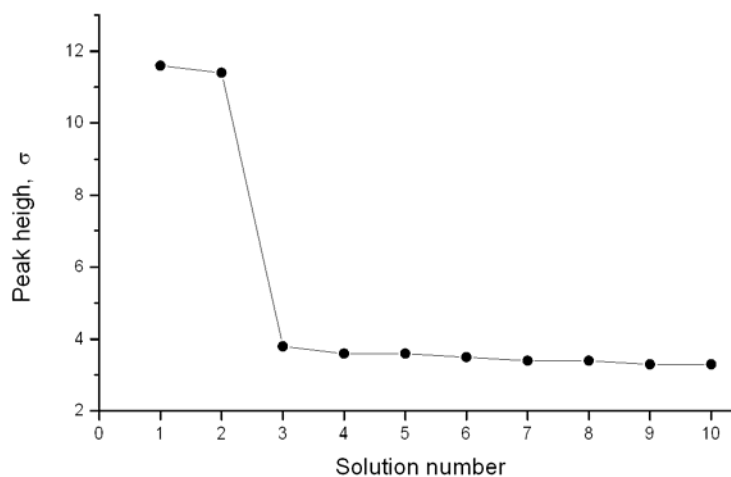As can be seen from Tables 6.9-6.11 and Fig. 6.2 there are always two peaks standing out.

Figure 6.2: Rotation function solution spectrum.

These represent two possible solutions, which have approximately equal values of α and β, but their γ angles differ by 180 degrees. It will be shown in Chapter 6.5 that these two solutions reflect the non-crystallographic symmetry of the tetramer molecule. Different combinations of resolution limits and integration radii were tested (Tables 6.10-6.11). In all cases the solutions are clearly observable over all the other background solution. These results show that the solutions are consistent and most probably correct, and that the rotation function can be solved for multilayered crystals.

6.3.3 The Translation Function Search

The top peak in the list of the rotation function solutions, produced by *AMoRe*, is not necessarily the correct one. If the rotation function solution is correct, it has to give the highest peak in the translation function as well. Since we know the orientation and the position of the search probe within the crystal unit cell, we can compute the structure factors from this model. Thus, we can judge the correctness of the solution by looking at the correlation coefficient and R-factor computed between calculated and observed structure factor amplitudes (Chapter 5). In the case of a correct solution, the correlation coefficient should be as high as possible and the R-factor as low as possible.

A translation function can be computed, within the *AMoRe* package, by means of the TRAING program (Fig. 6.1). It uses, as input information, the solutions of the rotation function. Different resolution limits can be tried (Table 6.13). Several rotation solutions for one search probe can be tested easily. For each rotation function solution the search probe is rotated by the specified angles and translated within the unit cell.

The program TRAING also computes the correlation coefficient (CC) and the standard crystallographic R-factor for each translation solution (equation 5.15). These two indicators are usually used to assess the agreement between observed and calculated structure factor amplitudes. The output of the translation function search contains the peaks sorted by the ratio between the correlation coefficient and the R-factor.

6.3.3.1 Results of the Translation Search

To calculate a translation function, the first 20 peaks of the rotation function were used as input. The different combinations of the resolution ranges were used for different intervals of 20 to 4 Å. The height of the produced peaks was limited to half the height of the maximum peak. The output list (Table 6.12, Fig. 6.3) contained only two strong peaks which had a correlation coefficient of 28.8 (26.9 %) and a R-factor of 38.8 (39.1 %). These two peaks are stable against different resolution limits (Table 6.13). The correctness of the solutions is corroborated by the fact that the highest peaks of the rotation function are the top translation function peaks. Furthermore, they have high correlation coefficients and there is a big difference between the two top solutions and background noise.

The results of the translation search show that a structure can be solved successfully by molecular replacement using electron diffraction data from multilayered crystals of catalase. It means that we are able to find the unknown orientation and position of a molecule of known structure within the unit cell.

Table 6.12: The list of the first 10 peaks translation function (the two correct solutions are highlighted).

Peak Number	α	β	γ	T_x	T_y	T_z	CC	R-factor
1	20.57	105.04	188.04	0.4567	0.1135	0.1969	28.8	38.8
2	20.70	105.12	8.04	0.4573	0.1133	0.1969	26.9	39.1
3	18.57	107.44	179.72	0.2770	0.4702	0.2893	9.0	41.9
4	17.91	107.36	359.03	0.3028	0.4161	0.3583	7.8	41.3
5	30.32	114.18	215.75	0.0532	0.1766	0.2889	7.8	41.9
6	17.77	72.74	5.87	0.0564	0.0847	0.4088	7.7	42.0
7	28.38	108.13	212.47	0.2441	0.3072	0.0455	6.7	41.9
8	164.49	53.43	325.64	0.0120	0.3796	0.2457	6.2	41.0
9	24.75	104.11	26.78	0.4439	0.3332	0.3158	6.1	42.0
10	15.51	126.57	145.64	0.4902	0.3773	0.2574	5.9	41.0

Resolution limits: 8 - 4 Å were used for the calculation

Table 6.13: The effect of the resolution limits on the correct solutions of the translation function (the best combination of resolution limits is highlighted).

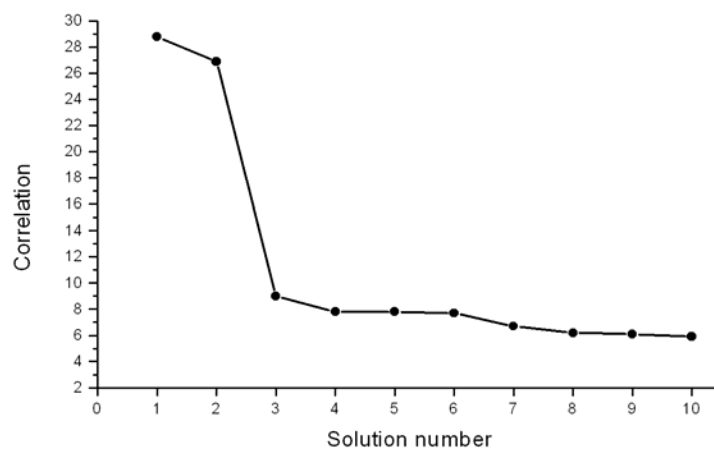| Resolution Limits, Å	Solution #1, CC	Solution #2, CC	The highest background peak, CC	Solution #1/noise	Solution #2/noise
20-4	53.0	52.2	38.0	1.4	1.4
18-4	51.3	50.5	35.8	1.4	1.4
16-4	49.3	48.7	32.2	1.5	1.5
14-4	46.5	45.7	28.7	1.6	1.6
12-4	43.7	42.5	24.3	1.8	1.7
10-4	34.2	32.4	13.3	2.6	2.4
8-4	**28.8**	**26.9**	**9.0**	**3.2**	**3.0**

Figure 6.3: Translation function solution spectrum.

6.4 Rigid-body Refinement of the Molecular Replacement Solutions

A very powerful component of *AMoRe* is its fast rigid-body refinement procedure, which is very efficient in the evaluation of the correctness of the molecular replacement solution. The ideas on which it is based were first developed by Huber and Schneider (*84*). It is not a least-squares rigid-body refinement of coordinates, but rather a minimisation of the misfit, where rotations and translations are calculated by interpolation and phase shifts.

After finding the solutions of the rotation and translation functions, the rotation angles and translation parameters are usually refined in *AMoRe* to minimise the quadratic misfit described by the following formula:

$$Q(k, B, \alpha, \beta, \gamma, x, y, z) = \sum_{\mathbf{h}} \left(\left| F_{obs}(\mathbf{h}) \right| e^{-B|\mathbf{h}|^2} - k \left| F_{\alpha,\beta,\gamma,x,y,z}^{calc}(\mathbf{h}) \right| \right)^2 \qquad (6.2)$$

Where k is a scale factor and B is the overall temperature factor. The refinement algorithm is performed in *AMoRe* by the program FITTING. This rigid-body refinement is considered to be another checking procedure, to prove the correctness of the solutions. If the solutions are correct, the refinement should result in an increase of the correlation coefficient and a decrease of the R-factor.

A rigid-body refinement of the catalase tetramer molecular replacement model yielded the final correlation coefficient of 28.8 (27.0) % and a R-factor of 38.2 (38.8) % between the observed and the calculated structure factor amplitudes of the two top solutions (Table 6.14). The values of R-factors are lower than the initial ones, suggesting that that these solutions are the correct ones.

Table 6.14: The list of the first 10 peaks translation function after rigid-body refinement (the two correct solutions are highlighted).

Peak Number	α	β	γ	T_x	T_y	T_z	CC	R-factor
1	20.85	105.24	188.17	0.4566	0.1103	0.1966	28.8	38.2
2	20.83	105.37	7.90	0.4540	0.1153	0.1971	27.0	38.8
3	18.24	107.78	180.29	0.2689	0.4680	0.2911	9.2	41.1
4	18.00	107.98	359.09	0.3029	0.4167	0.3557	7.6	41.1
5	31.34	114.33	216.56	0.0530	0.1767	0.2877	7.6	41.5
6	18.57	72.15	4.48	0.0583	0.0850	0.4074	7.4	41.4
7	29.35	108.41	212.47	0.2419	0.3075	0.0456	6.4	41.4
8	15.64	127.08	146.21	0.4888	0.3776	0.2585	6.0	40.8
9	164.81	53.23	325.12	0.0122	0.3794	0.2483	5.7	41.0
10	25.01	103.68	27.00	0.4424	0.3332	0.3199	5.3	41.5

Resolution limits: 8 - 4 Å were used for the calculation

6.5 Analysis of Two Top Solutions

The tetramer probe molecule was rotated and shifted according to the obtained two top molecular replacement solutions, using the program PDBSET (CCP4). The relative location of these models was analysed using graphical software, such as the program O (*85*). Careful analysis showed that these two molecules occupy the same space in the unit cell and relate to each other by non-crystallographic two-fold symmetry. Since the catalase molecule is a dimer of dimers, there are two ways to build the tetramer. Schematic representation of these two orientations may be depicted as:

A (C')	B(D')
D (B')	C(A')

Figure 6.4: Superposition of the monomer C (solution #1) and the monomer A (solution #2). Green: monomer A, solution2; yellow monomer C, solution 1.

Where A, B, C, D are monomers in the first orientation; A', B', C', D' are monomers in the second orientation. As can be seen from Figure 6.4, that the difference between solution #1 and local symmetry related solution #2 is — as expected — small. It means, that only one molecular replacement solution must be analysed.

6.6 Analysis of the Crystal Packing

An additional verification of the correctness of the obtained solution is an analysis of the molecular packing in the crystal. Since it is physically impossible for the molecules to overlap in three-dimensional space, the molecule must pack without overlapping with one of its neighbours. In some cases, the search structure may have a loop that is not in the unknown structure and, thus, a false overlap may be found. Another reason for the false overlap is the slightly different orientation of protein domains due to crystal lattice forces. Nevertheless, the amount of the overlap must be limited and a correct molecular replacement solution must have no or small clashes between symmetry related molecules.

Analysis of the new determined molecular packing for the crystal form II (our catalase crystals used for EC) showed that molecules are evenly distributed through the whole unit cell. No overlap between molecules and close contacts were observed. The crystal packing within the unit cell for two crystal forms is shown in figure 6.3.

Figure 6.6 shows a crystal packing of the two different crystal forms. In the crystal form I tetramer molecules are tightly packed in such manner that "square-like" channels are formed (Fig. 6.6a). Analysis of protein-protein contacts between symmetry related molecules reveals that each monomer has two hydrophobic patches on the surface in the regions of residues 100-105 (patch1) and 445-450 (patch2). The tetramer molecules form the following contacts: A_{patch1}-D'_{patch1}; A_{patch2}-D'_{patch2}; B_{patch1}-C'_{patch1}; B_{patch2}-C'_{patch2}; C_{patch1}- B'_{patch1}; C_{patch2}- B'_{patch2}; D_{patch1}- A'_{patch1}; D_{patch2}- A'_{patch2}, where A, B, C, D are monomers of the catalase tetramer, A', B', C' and D'- symmetry related monomers.

In contrast to the crystal form I, the monomer molecules in the crystal form II have two hydrophobic patches in the region 280-290 and 482-490 which form protein-protein contacts between A-C' and D-B' monomers. As result of these interactions, tetramer molecules do not form tight protein-protein contacts. This observation may explain the fact that all attempts to grow large crystals (>10 μm) have been failed. The tetramer molecules are uniformly distributed in the space forming protein layers (Fig. 6.6b) which are one half unit vector **c** thick. The latter is in agreement with studies of catalase crystal thickness distribution (*86*). This work showed that catalase crystal thickness can be expressed as:

$$d = \frac{c}{2}n \qquad\qquad (6.3)$$

where d is crystal thickness, **c**- an unit cell vector, n- a natural number.
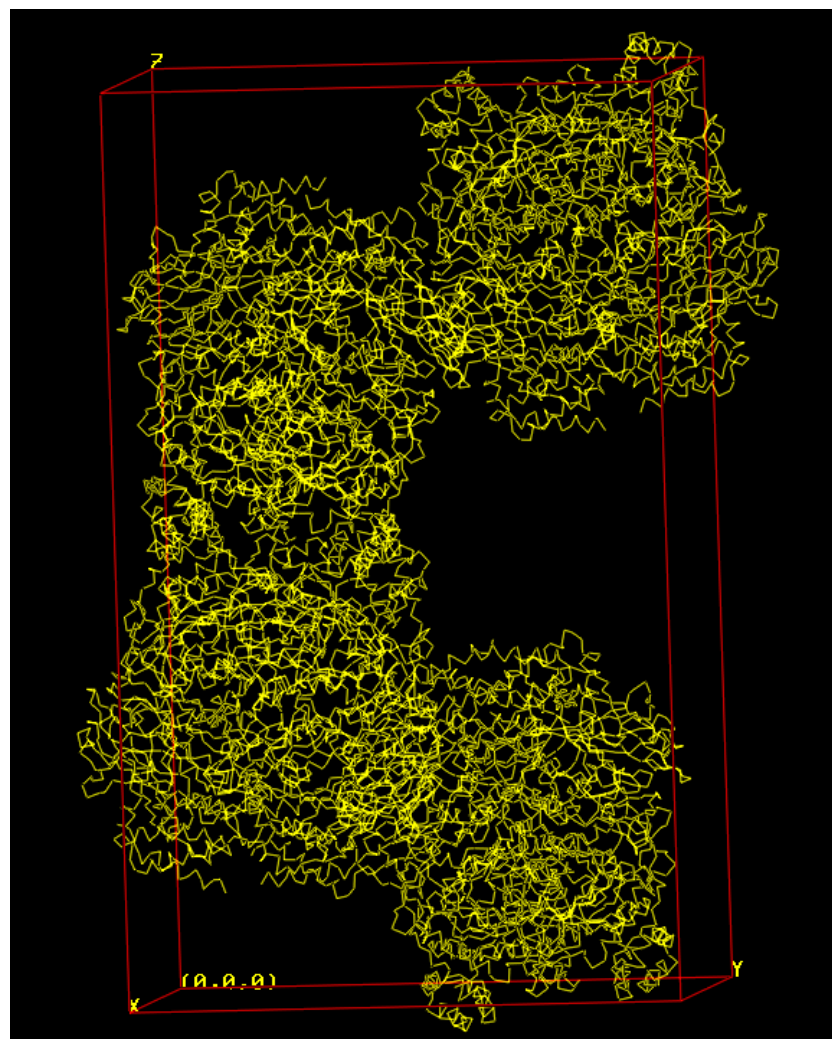
A
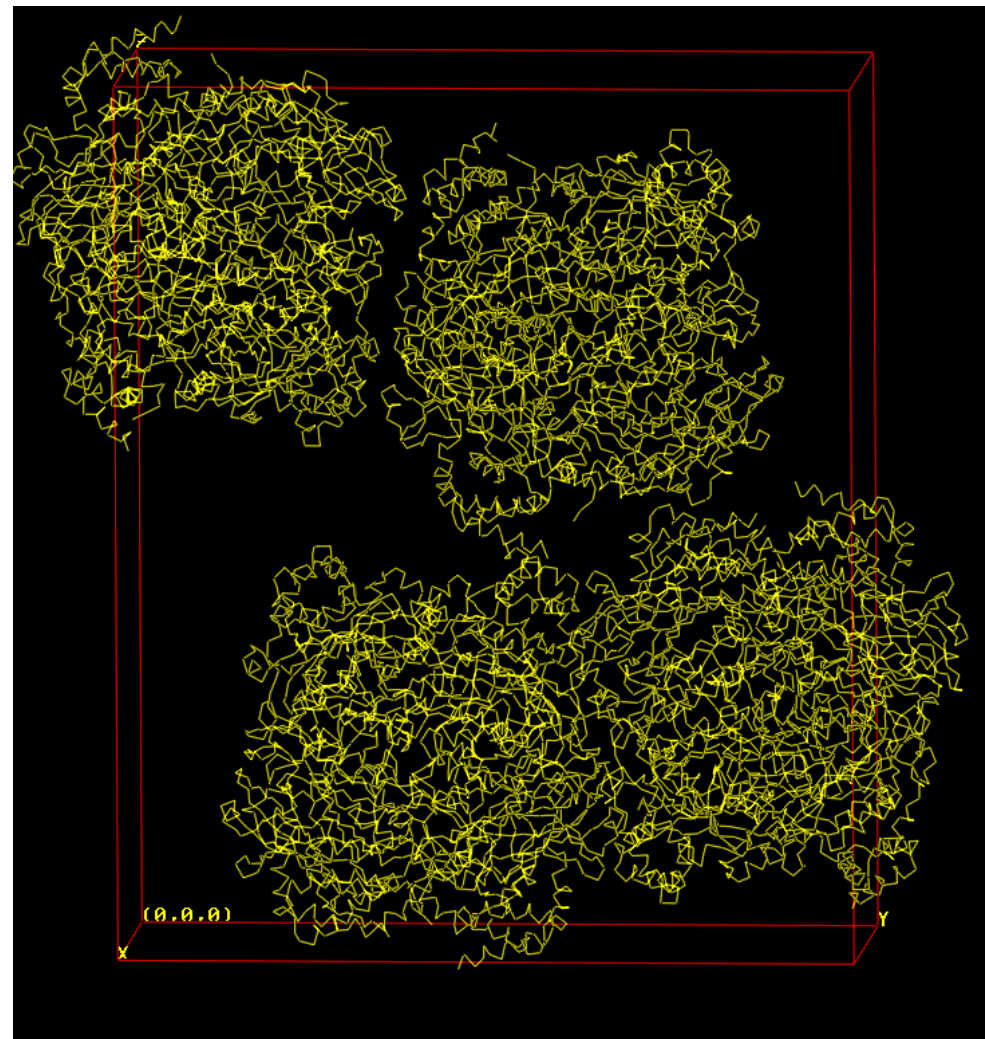
B

Figure 6.5: Crystal packing for the two different crystal forms in the unit cell. The 4 tetramers are shown in the unit cell. a) Crystal form I: a=87.8 Å, b=140.6 Å, c=232.4 Å , Space group $P2_12_12_1$; b) Crystal form II: a=69 Å, b=173 Å, c=206 Å , Space group $P2_12_12_1$.
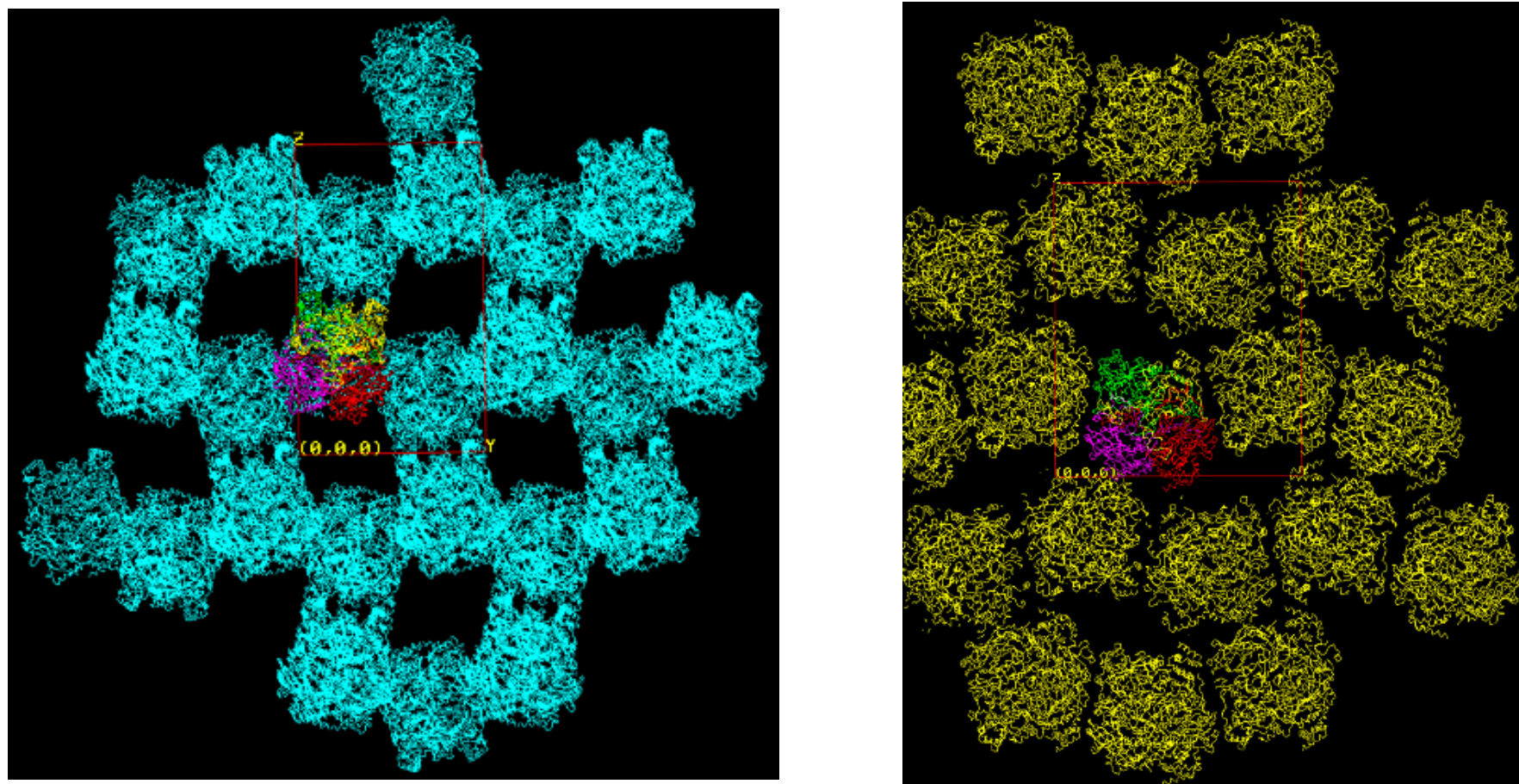
Figure 6.6: Crystal packing for a) crystal form I and b) crystal form II. A structural element – a tetramer molecule - is encoded by different colours (green, yellow, red and magenta).

6.7 Non-crystallographic Symmetry and Self-rotation Function

The self-rotation function for the crystal form II (EM diffraction data) was calculated using POLARRFN (CCP4) (*59*). It calculates the self-rotation function in polar angles (see Chapter 5). If there is a non-crystallographic two-fold or two-fold screw axis, prominent peaks must be found on the $\chi=180°$. A stereographic net produced by POLRARFN is used to measure ϕ and ω angles between points on a stereographic projection plot.

Since the molecular replacement solutions using a catalase tetramer molecule (two dimers) was found, the self-rotation function must show two strong peaks on the $\chi=180°$ section corresponding to the non-crystallographic twofold axes. However, no non-crystallographic peaks were found at $\chi=180°$ for the EC diffraction data (see Fig. 6.6). There are four possible explanations of these findings, which we will test and discuss in the following:

1) Low data quality. Diffraction spots could be affected by the dynamical scattering effect that would result in a noisy Patterson map.
2) Incomplete data (missing spots at random positions).
3) Missing cone effect. Since diffraction pattern at high tilt angles could not be collected, there is anisotropy (lack of data) in the Z-direction.
4) The NCS axis is parallel to the crystallographic axis.

To investigate all four effects, the following strategy was used. Since the orientations of the molecules for the crystal form II (catalase crystals used for EC) was known after solving molecular replacement, the structure factors were calculated from the model by the CNS (*60*) module "model_fcalc". Using calculated structure factor diminishes the break of Friedel symmetry due to the dynamic scattering effect and increase data quality statistics since possible experimental errors such as intensity integration problems due to our strong background gradient (Fig. 4.5) are not included. To study the effect of incompleteness and anisotropy in one direction, 12 datasets with calculated structure factors were produced:

1) 100isotropy. All calculated structure factor were included. The completeness of the dataset was 100%.

2) 90isotropy. This isotropic dataset contained calculated structure factors equally distributed in the space, but 10% of the spots from 100isotropy were randomly removed.

3) 80isotropy. This isotropic dataset contained calculated structure factors equally distributed in the space, but 20% of the spots from 100isotropy were randomly removed.

4) 70isotropy. This isotropic dataset contained calculated structure factors equally distributed in the space, but 30% of the spots from 100isotropy were randomly removed.

5) 60isotropy. This isotropic dataset contained calculated structure factors equally distributed in the space, but 40% of the spots from 100isotropy were randomly removed.

6) 50isotropy. This isotropic dataset contained calculated structure factors equally distributed in the space, but 50% of the spots from 100isotropy were randomly removed.

7) 40isotropy. This isotropic dataset contained calculated structure factors equally distributed in the space, but 60% of the spots from 100isotropy were randomly removed.

8) 10missing_cone. This anisotropic dataset contained calculated structure factors from 100iso with removing all spots which belonged to a cone oriented in the Z-direction with vertex of the cone placed at (0, 0, 0) and a cone angle equal to 10°.

9) 20missing_cone. This anisotropic dataset contained calculated structure factors from 100iso with removing all spots which belonged to a cone oriented in the Z-direction with vertex of the cone placed at (0, 0, 0) and a cone angle equal to 20°.

10) 30missing_cone. This anisotropic dataset contained calculated structure factors from 100iso with removing all spots which belonged to a cone oriented in the Z-direction with vertex of the cone placed at (0, 0, 0) and a cone angle equal to 30°.

11) 40missing_cone. This anisotropic dataset contained calculated structure factors from 100iso with removing all spots which belonged to a cone oriented in the Z-direction with vertex of the cone placed at (0, 0, 0) and a cone angle equals to 40°.

12) 50missing_cone. This anisotropic dataset contained calculated structure factors from 100iso with removing all spots which belonged to a cone oriented in the Z-direction with vertex of the cone placed at (0, 0, 0) and a cone angle equals to 50°.

Figures 6.7 – 6.9 show the results of calculations of the self-rotation function for the 12 datasets. As can be seen from Fig. 6.7, 6.8, for all isotropic data with completeness 40-100% the two peaks (arrows) corresponding to the two non-crystallographic axes of the two dimers can be recognised easily. It means that the low completeness is not limiting for finding of local symmetry in case of isotropic data. However, data anisotropy influences the self-rotation function significantly (Figure 6.9). If the missing cone is equal or more than 30 degrees, there are no significant peaks on the $\chi=180°$ section. Although the average completeness for the dataset with 30° missing cone is about 87 %, the anisotropy effect makes the self-rotation function uninterpretable. Since all EM data have missing cone of at least 30°, the self-rotation function calculation has to be interpreted with great care. Even if there is no direct evidence of local symmetry, all possible models (monomers, dimmers etc) should be tested to solve molecular replacement.
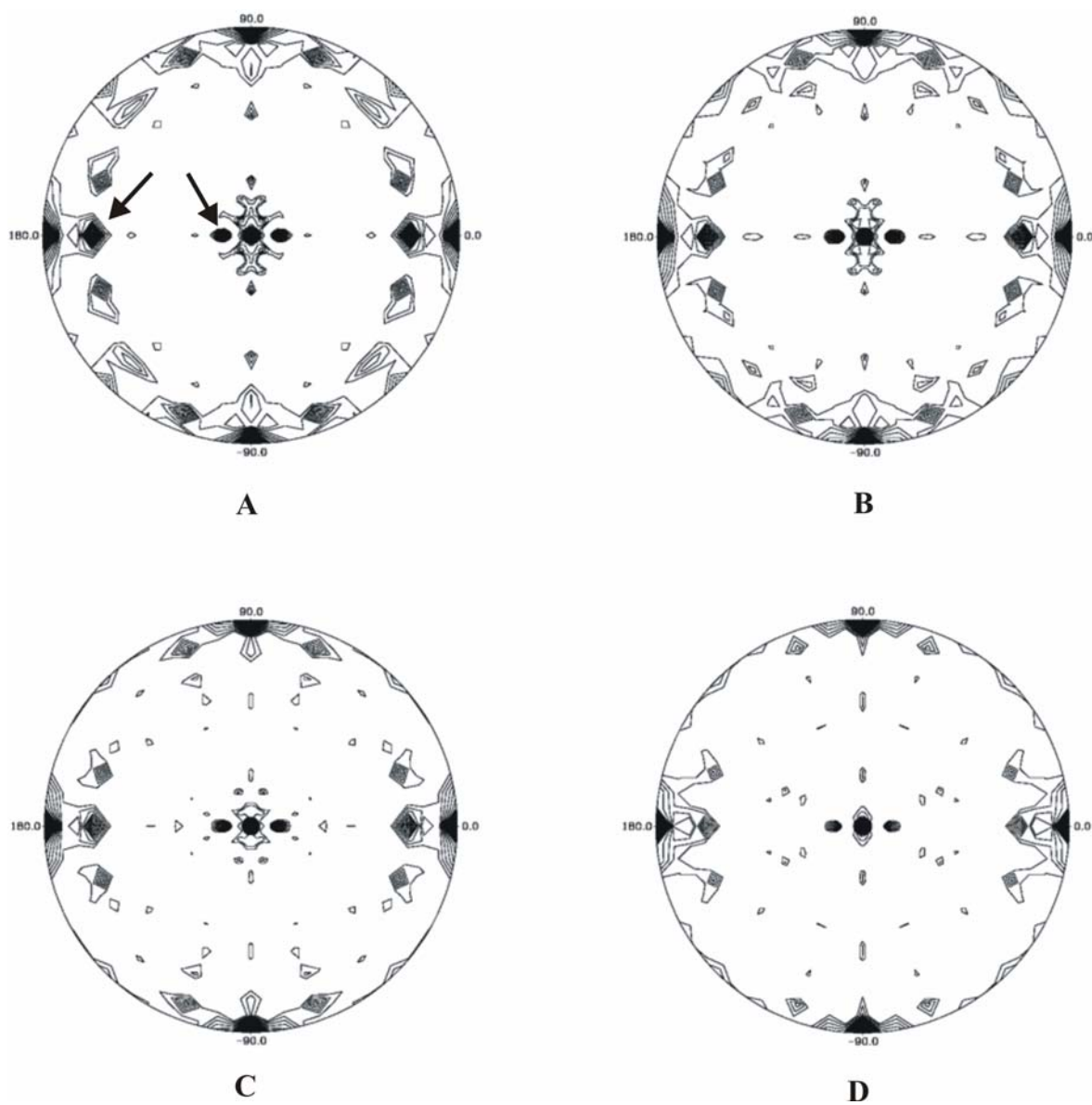
Figure 6.7: Self rotation function. A-100iso, b-90 iso, c-80iso, d-70iso. Resolution 10-4 Å, Patterson radius 40 Å.
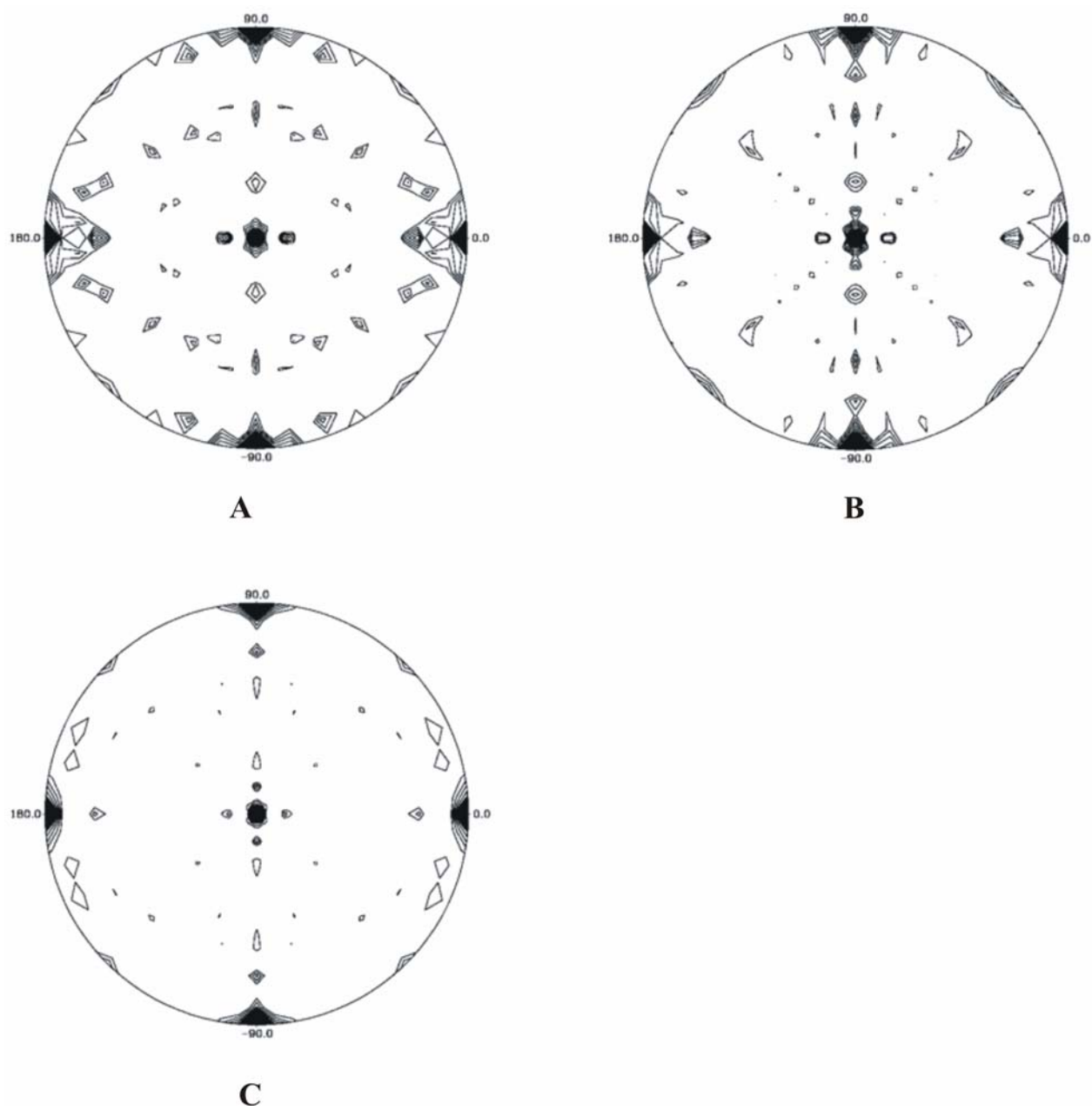
Figure 6.8: Self rotation data. a-60iso, b-50 iso, c-40 iso. Resolution 10-4 Å, Patterson radius 40 Å.
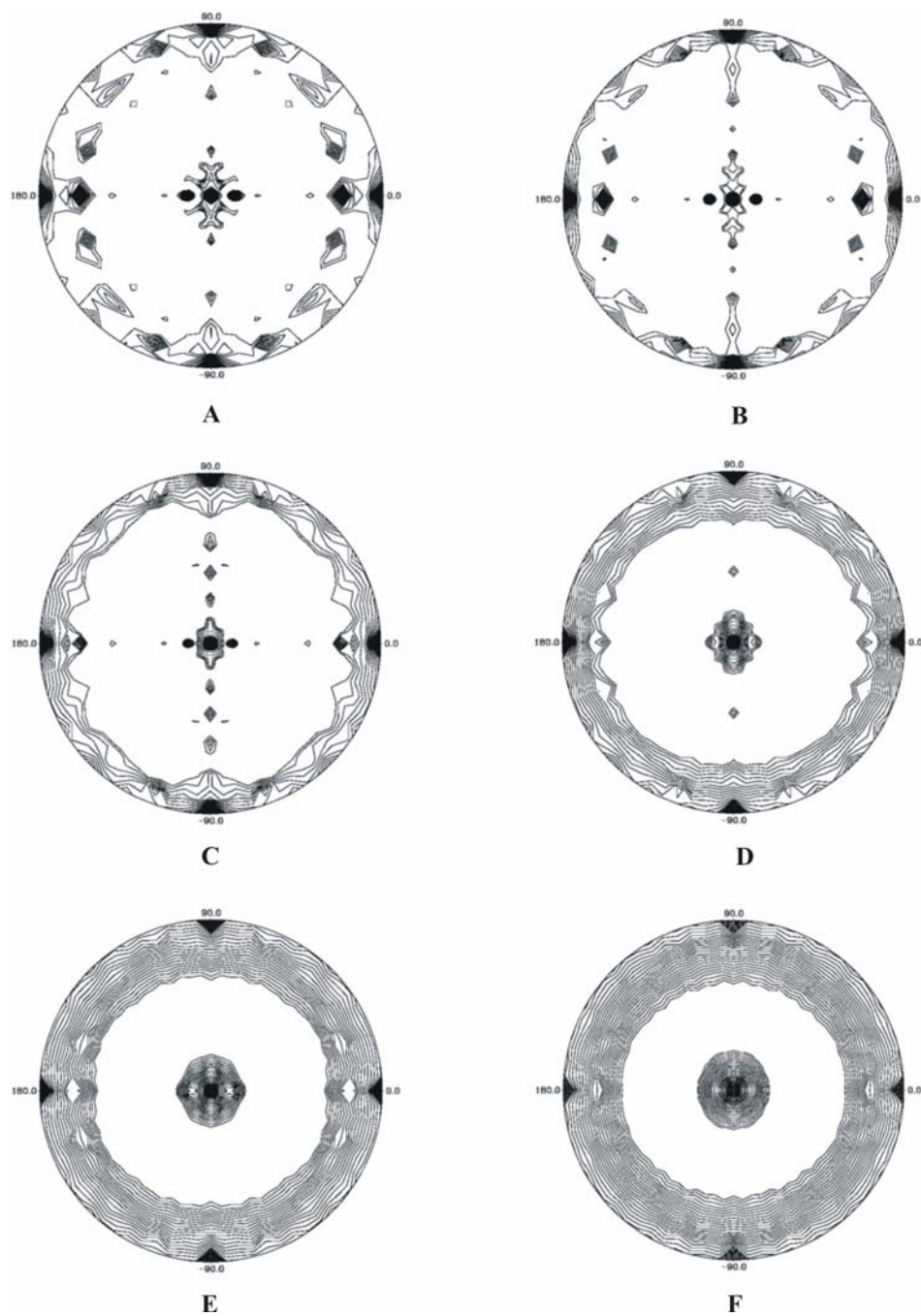
Figure 6.9: Self rotation function. Missing cone effect. Missing cone A- 0°, b-10°, c-20°, d-30°, e-40°, f-50. Resolution 10-4 Å, Patterson radius 40 Å.

6.8 Refinement of the Catalase Structure and Map Calculation

After the position of an initial model is determined the structure needs to be refined to obtain a set of atomic coordinates that corresponds best to the observed data. The initial model is only an approximation, and the structure factors calculated from the model may be in poor agreement with the observed structure factors. To do structure refinement, i.e. to minimize differences between calculated (F_c) and observed F_o structure amplitudes, visual rebuilding of the protein model is used. For this task different types of Fourier maps must be calculated and interpreted:

1) F_o-F_c, or Difference map. In this map there are peaks where density is not accounted for in the model to calculate F_c and holes where there is too much density in the model. This map is especially useful for finding missing water molecules, movements of side chains, movements of loops, etc.

2) $2F_o$-F_c map. This map is easily to interpret since it looks like protein density. However, the $2F_o$-F_c map is very sensitive to model bias.

3) Omit map. In the omit map the portion of the model to be examined is left out of the phase calculation, and the rest of the model is used to calculate phases of this portion of the map. It is possible because all parts of the model contribute to every diffraction reflection. The omit map shows much lower model bias than the $2F_o$-F_c map and, therefore, can be used to judge quality of the protein model.

When molecular replacement solutions were found using diffraction data from multilayered catalase crystals, the composite omit map was calculated. Figures 6.10 and 6.11 show examples of fitting the protein model into the density. Careful analysis of the omit map reveals that the map is of poor quality and subsequent model building and refinement are not possible. Some regions of the map (Fig 6.10) have very nice agreement between an amino acid chain and density. However, others have disconnectivity along the protein chain (Fig. 6.11). The main common feature of the omit map is that it has elongated blobs of density in the Z direction (Fig. 6.12). We conclude that the reason of such elongation is the missing cone of data along the axis.

To analyse further the influence of the missing cone effect on map interpretation, we calculated omit maps using datasets 10missing_cone-50missing_cone (Chapter 6.7). It was shown that if a missing cone is larger than 40° (75 % completeness), it is hard to interpret maps correctly. In the case of our current catalase dataset the completeness is 50-70% for different resolution shells and, therefore, protein model building and improvement is not possible.

Nevertheless, with innovations discussed in the next chapter a dataset with sufficient completeness (up to 87 %, see Chapter 4) can be collected, and future studies should analyse the possibility of model refinement for multilayered protein crystals using, for example, various correlation techniques..
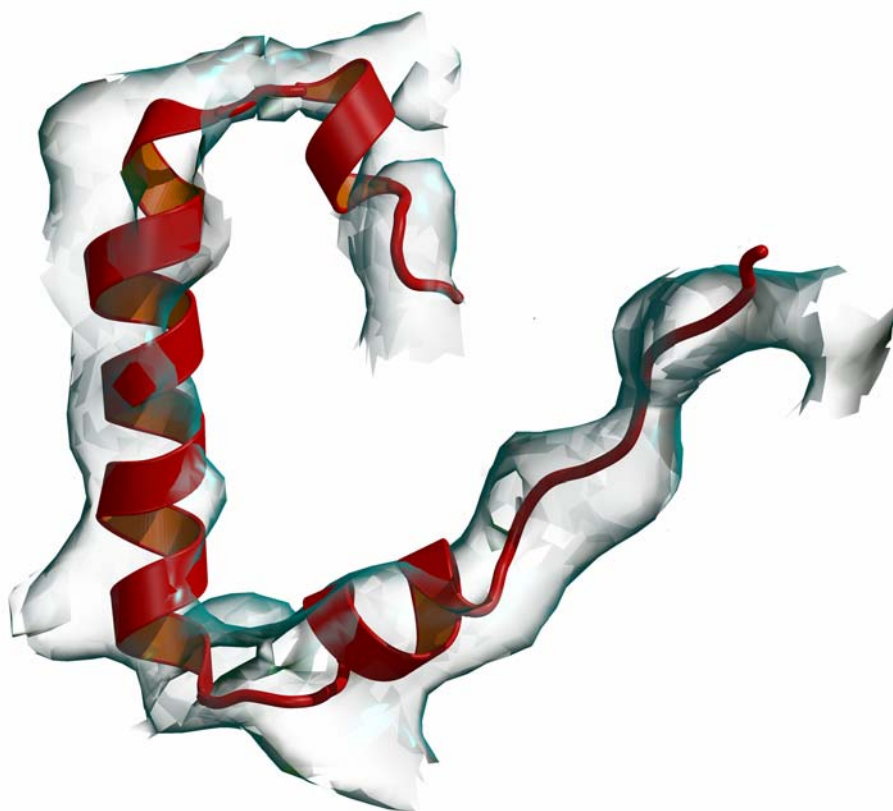
Figure 6.10: Example for fit of an amino acid chain (ribbon representation) to electron density map (cyan) at 4 Å resolution.

Figure 6.11: Fit of the molecular model (yellow) after rigid body refinement into a $2F_o - F_c$ composite omit map (cyan) at 6 Å resolution.

Figure 6.12: Effect of the missing cone on reconstructed map interpretation. The systematically missing data in the direction of the lattice vector **c** leads to elongation of nucleus density in the Z direction. This results in complication of model rebuilding and map interpretation.

Chapter 7

Discussion

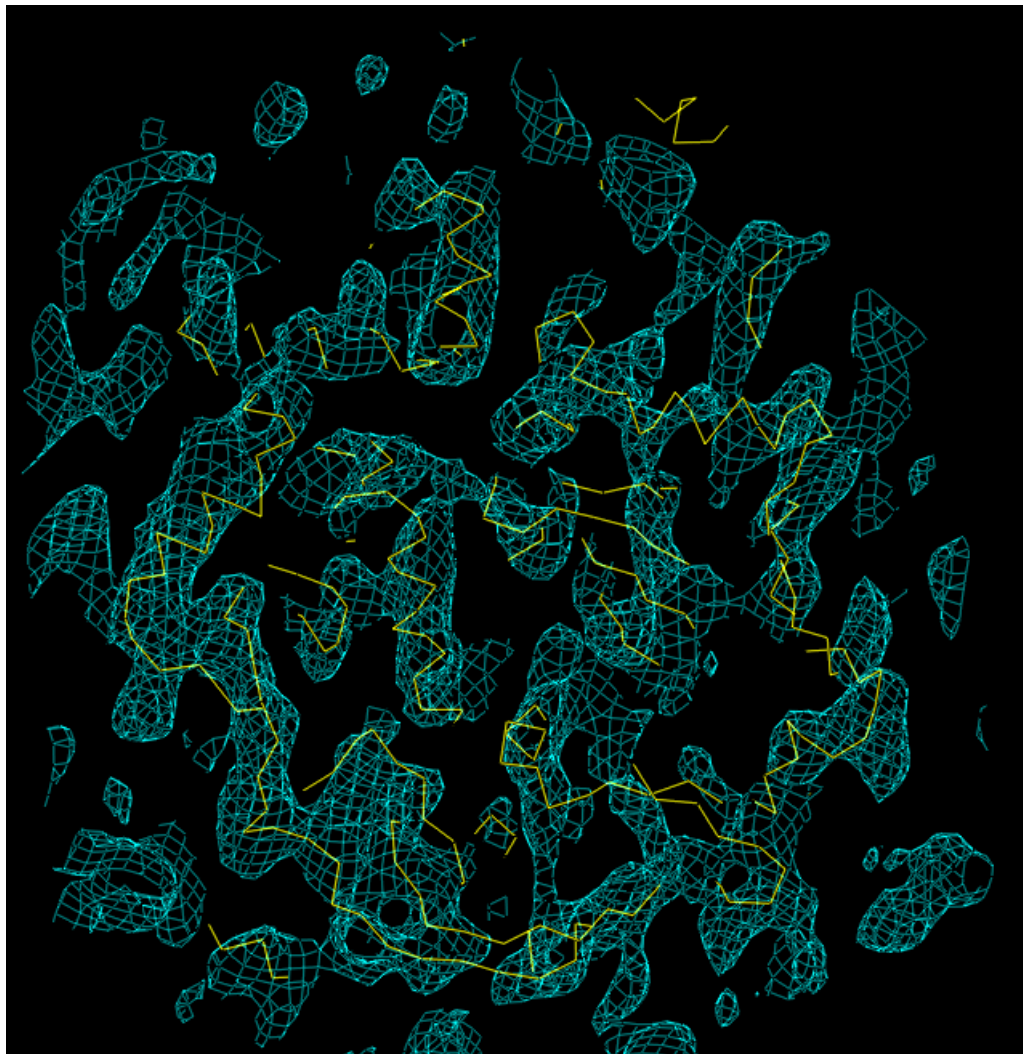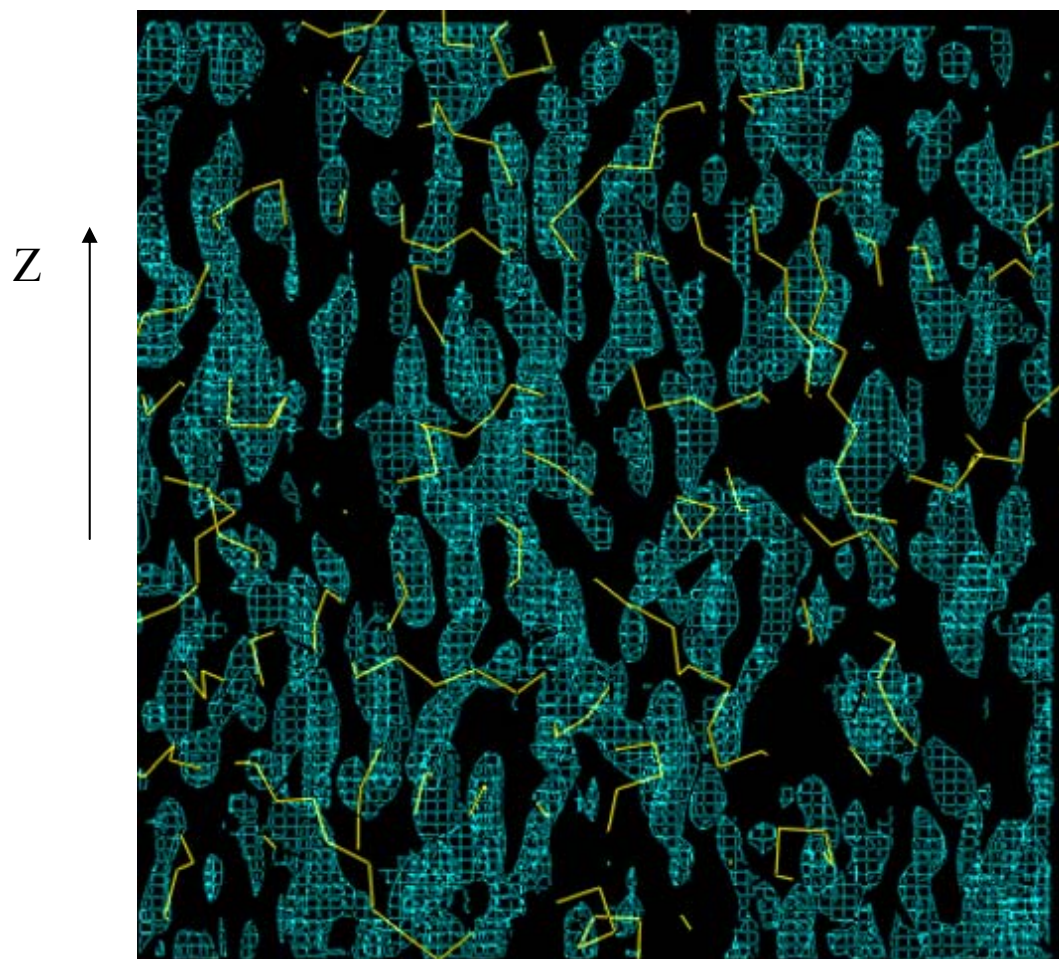The aim of this project was the experimental proof of principle of electron crystallography on "multilayered crystals"[16] to obtain high resolution structures of biological macromolecules and protein complexes. This thesis has completed the methodological work initiated by Burmester, Dimmeler and Vossen (*22-24;86-88*). Their studies concentrated on several details of the diffraction data collection and processing. It showed the theoretical possibility to extract structural information from diffraction data of multilayered crystals. However, once the work by Burmester, Dimmeler and Vossen was completed it became clear that this method had to be tested using actual experimental data and a real crystal structure needed to be determined.

The achievement of this thesis is the first determination of a protein structure (bovine liver catalase) using electron diffraction of multilayered crystals. To obtain this result, four tasks had to be solved:

1) The recording of a complete 3D diffraction data set;
2) Indexing of diffraction spots and intensity integration;
3) Solving the phase problem by Molecular Replacement;
4) Density reconstruction and model refinement.

With a real 3D experimental data set at hand we analyzed also necessary modifications and extensions for the usual X-ray processing methods when applied to Electron Crystallography.

---

[16] Multilayered crystals are thin three-dimensional microcrystals which have limited periodicity (usually only a few layers) in one direction.

7.1 Recording Complete 3D Data Set

The first step was the recording of high quality diffraction patterns from differently tilted multilayered crystals. The quality of the diffraction patterns must be good enough to use them for quantitative interpretation. Usually, the recording of several diffraction patterns with high signal-to-noise ratio from one crystal is an impossible task because of strong radiation damage of the sample and an inelastic background problem, which affects the recorded intensities. For example, several groups have tried to collect a complete dataset from multilayered crystals but neither density nor model structure was reported[17].

Bovine liver catalase was chosen as a model protein for our studies. It forms platelet-like multilayered crystals (see Fig 4.4), which have been used intensively in EM studies (*89*). Although the catalase crystals diffract to resolution higher than 3Å, the recording of a complete 3D data set at high resolution and good quality needed modified experimental setups and a highly trained microscopist.

The set of high resolution catalase diffraction patterns was recorded at different tilt angles, while the frozen hydrated specimen was kept at cryogenic temperature (100 K). Radiation damage is less severe at low temperatures than at room temperature by a factor of three to six. Furthermore, embedding in vitreous ice minimizes sample damage during preparation steps and helps to obtain high-resolution data.

The use of modern techniques such as sensitive detectors (Imaging Plates with high dynamic range) and removal of the inelastic background by energy filtering yields accurate data that can be used for data processing.

The energy source must be chosen with great caution. A low energy source increases the contribution of dynamic scattering (figure 2.7), which reduces the information content and interpretability of the data. Unfortunately, the data used here had to be collected at an electron energy of only 120 kV. However, using low dose conditions (2.5 $e^-/Å^2$) allowed to record up to 5 diffraction patterns from one protein crystals. A typical number of unit cells

---

[17] Personal communications of groups from the United States and Japan at conferences.

within the illuminated area was 500.000 (for 5 unit cell layers), i.e. each diffraction pattern results from diffraction of about 8 million single molecules.

To collect a dataset with maximal possible completeness, it is necessary to sample the maximal volume of Fourier space, so high maximum tilt angles were required. Experimentally, a modern cryoholder allows to collect data with a tilt angle upto 60°. The incompleteness of data is known as "missing wedge" effect. However, recording of diffraction patterns from highly tilted crystals is a real challenge. Indeed, the visible section of the electron microscopy grid diminishes and only a few grid squares are visible when the tilt increases. Consequently, the probability of having catalase crystals at the right place on the grid is very low.

Nevertheless, we were able to record 55 diffraction patterns of 17 tilt   series (i.e. one crystal per tilt series) with tilt angles from 0° to 50°. Data from these diffraction patterns have been used for processing and reconstruction. Data completeness is in the range of 63-74% (10-20 Å) and 40-63% (4-10Å). It is known that lower completeness results in noisy elongated electron density maps which makes it difficult to interpret the maps and to built an atomic model. We showed that it is theoretically possible to measure about 87 % of reciprocal space (for space group $P2_12_12_1$) of the used catalase crystals.

Numerical simulations were performed to analyse the missing wedge effect on density interpretation. It was shown in the case of catalase crystals that if data completeness is equal or higher 70 % the lack of information in one direction has a minor influence on electron density reconstruction (Chapter 4). It means that if the maximal possible complete data is recorded from catalase crystals with a modern cryoholder, data completeness is enough to perform full protein model refinement.

In general, as we discussed in chapter 4, the degree of incomplete data strongly depends on the symmetry of the crystal lattice and the geometry of the crystal. For example, for the cubic space group I432 or F432, the necessary tilt angle range would only be about 35° to collect complete data (Chapter 4).

7.2 Indexing of Diffraction Spots

The quantitative interpretation of electron diffraction patterns requires an exact determination of the diffraction geometry, described by Eulerian angles between the incident electron beam and the crystal axis. The determination is extremely important for all structural studies since many patterns from different tilt angles have to be merged and scaled to obtain the final 3D information (see Chapters 4.5, 4.6). Diffraction theory predicts that the position of each individual diffraction spot is analytically determined by the relative orientation of the diffraction beam and crystal lattice. It is thus possible to derive the Eulerian angles from a single diffraction pattern (*23*). To find the correct Eulerian angles, the spot positions must be determined accurately. Experimentally, the correct determination of spot positions depends on the mosaicity of crystal, divergence of beam, the point spread function and pixel size of the detector. All these parameters may result in broadening of the diffraction spot. In the case of EC the strong inhomogeneous inelastic backgrounds affects the spot position determination.

The method developed by Dimmeler was used to determine the orientation of the crystal lattice. The procedure utilizes all individual spot positions as input for a global non-linear least-squares procedure. This algorithm allowed precise determination of the orientation of diffraction patterns and merging the integral data into a complete 3D data set. Diffraction spots were indexed according to tilt geometry. The correct determination and indexing was extremely important because it has been shown that an incorrect indexing of the recorded diffraction intensities reduces the resolution of the final structure (*61*).

However, the method described above is not stable for low tilted diffraction patterns (<10°) and gives a large error for the tilt parameters. To determine the tilt geometry of low tilted crystals, another method, the so-called analysis of Laue zones, was used. This method utilizes information about positions of bands of reflections, which result from the intersection of the Ewald sphere with successive layers in 3D space. To verify calculated tilt angles and, therefore, correct index assignment, calculated tilt angles and the nominal increment of angles of one tilt series was used as an additional crosscheck. The differences between calculated tilt angles Θ (Dimmeler's method and Laue analysis) and the nominal

tilt increment between pairs of one tilt series were calculated and compared. The experimental data obtained in this thesis show that for low tilt geometry (0-10°), the tilt angle values determined by the Laue zone analysis are in better agreement with the nominal readings at the goniometer of the Electron microscope. However, the available number of diffraction patterns is not large enough for a statistical analysis to evaluate errors in tilt angle determination by these two methods. At this point we used both methods and discarded diffraction series with obvious disagreement between the two methods.

Once the positions and Miller indices for each diffraction spot were determined, the actual intensity of the spots must be measured. Since the structural information of the protein is encoded in the intensity distribution, an accurate intensity calculation is required. Unfortunately, the diffraction pattern includes sometimes a strong background. Information about background intensity, which results primarily from inelastically scattered electrons, and actual diffraction spot intensity must be separated. Therefore, we used an energy filter to remove inelastically scattered electrons and to decrease the background intensity. To record diffraction patterns at a quality sufficient for quantitative interpretation, we have used the energy filter with an energy band of 0-25 eV. Unfortunately, the separation of the strong background was not always possible. A residual inelastic background is present as the filter is not yet optimal. The background affects primarily innermost spots which carry information about low resolution features of the protein (> 30Å). Due to nonlinearity of the background, the center of mass of the spot is shifted, which makes it difficult to find the correct spot position. Although usually low resolution data is not included for the reconstruction of the electron density map, information about positions of the innermost spots is important for different steps of data processing such as azimuthal and radial distortion corrections (*21*) and for finding the correct tilt geometry. It means that in future using a better energy filter (with higher dispersion, see Appendix A) would be desirable to improve diffraction pattern quality and to increase signal-to-noise ratio.

7.3 Phase Problem

Determination of the phase is a problem inherent to crystallographic techniques. Diffraction data provide only "one half" of the information that is necessary for the reconstruction of the electron density map. This lack of information is known as the *"phase problem"*. In protein X-ray crystallography, the phase problem is currently solved by three methods: multiple isomorphous replacement (MIR), (multiple) wavelength anomalous dispersion (MAD, SAD, SIRAS), and molecular replacement (MR).

In conventional electron crystallography, imaging can be used for phase determination. Thus, for the reconstruction of 2D crystals, images of crystals were used to determine phases. In multilayered EC this is not possible[18]. An alternative method for phase determination has to be applied. One possible solution to the problem is the adaptation of X-ray crystallographic methods for phase determination to electron crystallography. MIR and MAD requires binding an atom of high atomic number to each protein molecule within the crystal. Then phase information is obtained by evaluation of differences in recorded intensities caused by the substituted heavy atoms.

Two studies investigated this effect on protein crystals. Ceska and Henderson (*90*), who tried to obtain phase information for bacteriorhodopsin crystals, came to the conclusion that in their case data of sufficient accuracy is not available. On the other hand Burmester and Schröder (*86*), analyzing catalase diffraction patterns with heavy atoms bound to the protein, found that significant intensity changes could be obtained and that it should be possible to use MIR and anomalous scattering of electrons for phase determination in electron diffraction. However, in any case, use of heavy atom derivatives for structure determination in EC will be more difficult than in X-ray crystallography. Burmester and Schröder showed that changes in intensities for MIR are only 4-6% in contrast to 10-17% in the case of X-ray (*86*). The weak expected differences should be detectable in electron diffraction patterns. To distinguish such differences, the determination of spot intensities,

---

[18] As has been studied by Vossen (*21*) the effect of dynamical scattering can only be "compensated" – or better - minimized by the processing of diffraction data. The images themselves are highly affected by dynamical scattering and only *a posteriori* simulation of a known structure can be used to evaluate experimental images. This is also well established for electron diffraction in material science.

and therefore the determination of the tilt geometry of the two crystals (native and heavy atom derivative) must be extremely accurate. However, at least at present, it does not seem to be possible due to experimental problems such as the residual background problem and the low S/N ratio at 120 kV electron energy (see discussion above). As a result, the differences in intensities will not only represent the presence of heavy atoms, but most likely also the recording of different parts of the same Gaussian intensity distribution in Fourier space (*22*). It would lead to phases with large errors and – most likely– an uninterpretable experimental electron density map.

The third method for solving a phase problem, the molecular replacement method, requires information about amplitudes and does not rely on experimental phases. In addition, it requires knowledge of a homologous protein structure to obtain the unknown orientation and the position of the molecule in the new crystal form. One can object that the molecular replacement method has significant limitations because it needs a homologous model, which may not always be available at present.

It was shown that proteins, which belong to one protein family, almost always have homologous structures (*91*). The search to determine the 3D structure of, at least, one protein from each family has been initiated a few years ago. In future this will mean that as soon as diffraction data are collected the "encyclopaedia of protein structures" should be searched to find a homologous structure for molecular replacement.

The other experimental situation where one does not necessary need *de novo* phasing is the structural study of protein-protein complexes. One can assume that structures of some individual components are known at high resolution. In this case molecular replacement is an indispensable method because their structures may be used as models for molecular replacement and to calculate initial phases. When the positions of the components are found, the structure of the unknown part is built by fitting a polypeptide chain into an electron density map. After fitting the model is refined to reveal small details, which could be important for understanding of biological functions.

As is clear from the discussion above the method of choice for solution of the phase problem in EC of multilayered crystals – at this point (see below) – is Molecular

Replacement. In this thesis we present the first protein crystal structure solved from EC diffraction data of multilayered crystals of bovine liver catalase by the molecular replacement method. As a homologous model the structure of the enzyme solved by X-ray crystallography was used (*58*). The phase problem was successfully solved. The rotation (CC=16.2 %, next peak CC=10.6 %) and translation (CC=53%, next peak CC=36.4%) functions show clear peaks corresponding to the correct orientation (for details see Chapter 6). This showed for the first time that electron diffraction data from multilayered protein crystals can be used to solve the phase problem by Molecular Replacement. These results open new possibilities and perspectives for structural biology.

7.4 ClpB

The true test of any structure determination method is of course a solution an unknown protein structure. This project was initiated at the time when a three-dimensional structure of the molecular chaperone ClpB was not known. Bacterial ClpB is an essential protein of the heat-shock response, which has the remarkable capacity to rescue stress-damaged proteins from an aggregate state (*92*). ClpB is a 95 kDa protein which forms a hexamer in solution. The protein was crystallized by Schlee *et al.* (*93*). The protein forms large
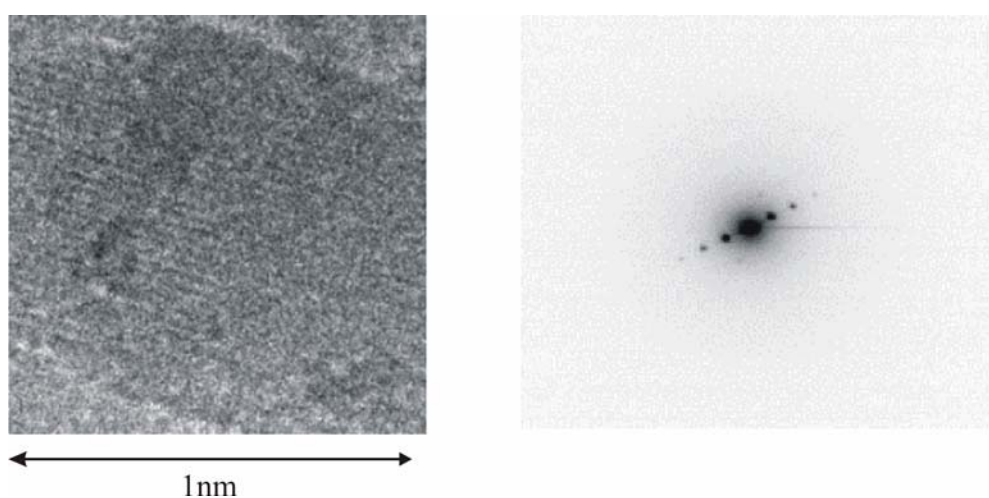

1nm

Figure 7.1: Electron micrograph of a ClpB crystal in vitreous ice (left) and diffraction pattern from the ClpB multilayered crystal recorded on an Imaging Plate.

crystals, which, however, are not suitable for X-ray analysis due to absence of diffraction power . It was suggested that electron diffraction is better for small multilayered crystals of ClpB since they may be more ordered internally. Preliminary crystallization experiments showed that small multilayered crystals of ClpB can be obtained under almost the same conditions which are required to grow large crystals. Fig. 7.1 shows electron micrographs of ClpB multilayered crystals that are embedded in vitreous ice. Indeed the diffraction pattern from the multilayered crystals can be recorded. However, high order diffraction spots were not observed. A possible reason is the small size of the crystal. A typical size of the catalase crystal (4-6 μm in one dimension), which diffracts to high resolution, is about 5-10 times larger than the investigated ClpB crystals. When small catalase crystals were tested for diffraction experiments, recorded diffraction patterns were similar in terms of resolution to the ClpB diffraction pattern depicted in Fig. 7.1. Therefore, we assume that when optimal crystallization conditions are found to produce relatively large multilayered crystals (3-5 μm), high-resolution diffraction patterns may be recorded from ClpB as well.

7.5 The Future of Electron Crystallography

In the near future multilayered EC should be one of the possible methods applied in structural biology studies. Using innovative technologies will allow to accelerate each step from crystallization to data collection by many times. Including techniques and strategies discussed below may increase the number of structures solved by multilayered EC dramatically. To reach this goal, future researchers have to focus on the following aspects of this method:

7.5.1 Crystallization

Usually researchers spend a long time in the laboratory to grow large, good diffracting protein crystals suitable for X-ray analysis. If small crystals are available, for example needles, attempts to get larger crystals often fail. Multilayered EC can utilize such small

crystals, for example, bovine liver catalase as shown in this thesis. As shown here, the methods used in X-ray crystallography can be applied directly to EC.

However, several modifications are required. Since protein crystals are extremely fragile objects, it is necessary to avoid various mechanical stresses during the preparation procedure. The most crucial step is crystal transfer.   If the protein crystallizes directly in the buffer solution, the crystals must be transferred onto the EM grid. Usually, using a pipette a drop of 3 to 5 µm is applied to the grid surface. One possible solution to decrease crystal damage is direct crystallization on the grid. If the EM grid is left in the solution over a longer period of time and the sitting drop geometry is used, crystals begin to grow directly on the surface of the grid or settle from the solution on the grid under gravitation forces.

Another problem is the large number of high-quality multilayered crystals needed for EM studies. Therefore, a high concentration of crystals on the grid will accelerate data collection. However, most crystals are lost when the excess buffer is removed from the EM grid by blotting with filter paper just before the sample is flash cooled in a mixture of liquid propane and ethane (see Appendix A). One possibility to increase the crystal concentration on the grid after blotting is using a sticky hydrophilic surface after crystallization. Small crystals will meet the surface due to diffusion motion and then they will be fixed on it. A sticky surface approach could be combined with centrifugation just before blotting. When extra gravitation force is applied, crystals could settle down from the solution on the grid. It is obvious from the discussion above that no conclusive theory for this crystallization problem exists and a lot of experiments will be necessary.

7.5.2 Sample Preparation

 The sample preparation rate can be increased by robotic techniques which are available commercially. Companies and researchers continue to develop more powerful microscopes

and sample preparation tools for use in cryoelectron microscopy. FEI's sample preparation robot (Vitrobot^{TM}) can prepare the sample in an almost fully automatic way. It plunge-freezes any aqueous suspension at optimal freezing conditions in liquid ethane or propane. The grid handling system performs the next step. The system automates the delicate task of transferring grids between the tray and the specimen holder with high accuracy. The robot then loads the specimen holder into the microscope and performs the various manipulations that would otherwise be carried out by the operator. All these procedure enable unattended, 24-hour microscope utilization.

7.5.3 Automatic Data Collection

The next big improvement will be using a fully automatic TEM. Automated methods for data collection and analysis will have a significant impact in transferring the cryoelectron microscopy technology to the general biological community as well as in increasing the volume of data that can be collected during a single session at the microscope. To achieve this goal, the special automated system that can emulate all of the decisions and actions of a highly trained microscopist in collecting data from a vitreous ice specimen must be applied. For example[19], Carragher *et al.* (*94*) have developed a system named Leginon to automatically acquire cryoelectron micrographs. This system performs in an automatic mode different operations such as identifying suitable areas of vitreous ice at low magnification, determining the presence and location of specimen on the grid, automatically adjusting imaging parameters (focus, astigmatism) under low-dose conditions, and acquiring images at high magnification to either film or a digital camera. This system can run unattended for over 24 hours and is responsible for every aspect of image acquisition. For example, a prototype of this system has obtained up to 1000 images per day of negatively stained catalase crystals (*95*). In contrast, a highly skilled microscopist would be able to record only about a few dozen images. We believe the Leginon system can be easily adopted for collection diffraction data from multilayered protein crystals.

---

[19] See also the newer issue of JSB, vol. 145, (2004) for more recent work.

One problem arising in the automated data collection system is automatic visualization and recognition of tiny crystals on the carbon film. X-ray protein microcrystallography also faced with the same problem. It has been solved by using diffraction based alignment. Furthermore, because of the small size of the crystals, it is extremely difficult to precisely align them. The requirement is that the microcrystal is in the microbeam and on the rotation axis. It requires an order of magnitude higher precision (that is, microns compared to tens of microns) in alignment than on standard setups. The radiation damage problem means that changing crystals and subsequent alignment need to be frequently repeated. The solution is to have completely automatic procedure for crystal and beam alignment. This system which use a completely automatic procedure for crystal and beam alignment has been developed and successfully applied for data collection at the ESRF (*96*).

In multilayered EC, diffraction based alignment cannot be used due to much stronger beam damage. Nevertheless, the problem of visualization can be solved in a very elegant way (for details see (*97*)). Since the scattering experiment takes place in an EM, a high resolution imaging device is integrated in the experiment. Switching to the image mode allows to visualize easily a crystal and align it on the optical axis.

Furthermore, an on-line evaluation of diffraction patterns and selecting good crystals before recording tilt series would be possible using hybrid pixel detectors (*98*). They were primarily developed for particle physics and could be an alternatice to CCD based detectors or Imaging Plates. The hybrid pixel detectors include many advantages such as immediate availability of data, better control over experiment, possibility of automating procedure, better a S/N ratio and a high dynamic range. Faruqi *et al.* (*98*) showed that preliminary experiments in recording 120 keV electrons using a pixel detector have produced extremely promising results. However, a number of problems such as a better estimate of resolution[20], the reduction in double counting of electrons in adjacent pixels and obtaining a larger sensitive area of the detector have to be solved in the near future. All these improvements will make a hybrid pixel detector more suited for electron microscopy.

---

[20] The relatively large size of pixel at present (~200-300 μm) is not problem, as the diffraction patterns can be magnified. The real problem is the number of independent pixels which should be in order of about 4×4 K to record pattern with typically 3000-4000 spots. In fact, depending on the physical process of electron detection, larger pixel sizes are desirable, as a higher dynamic range is realized for larger pixels. This is not important for low dose imaging but can be important for diffraction patterns.

7.5.4 Beam Damage

Radiation damage of the sample is a serious problem in EM. Nevertheless, we were able to record up to 5 high-resolution diffraction patterns from each catalase crystal. Radiation damage can be decreased further by using state-of-the-art modern microscopes. Firstly, liquid helium temperature should be used instead of liquid nitrogen temperature. Stahlberg H., Zemlin F. and Fujiyoshi Y.(*99-101*) have demonstrated that a further reduction of the specimen temperature from liquid nitrogen (~70K) to liquid helium (~4K) results in an improvement of the cryo-protection factor by a factor of two or more. Secondly, high-energy sources, for example 300 keV emission gun, can decrease radiation damage, as less inelastically dominated multiple scattering occurs(*56*) (Fig. 2.7).

7.5.5 Cryoholder

To accelerate data collection and decrease radiation damage, a new cryoholder system has to be used. Due to mechanical imperfections of the goniometer, the tilt axis is not completely stable. It leads to a sample shift when the grid is tilted. It means that particularly at high tilt angles (>40°), the microscopist needs to recenter and refocus the image after every tilt increment. This results in additional exposure, which destroys the sample, and is very time consuming. The same problem exists in cellular electron tomography. For these needs two groups have pioneered the development of dedicated instrumentation and methods for facilitating tomographic data collection and processing (*102*). These two systems can perform routine fully automatic data collection, including recording datasets, focusing and correction of image shifts. To adopt these systems for automatic data collection in the case of multilayered protein EC, modification of existing hard- and software will be required.

7.6 Conclusion

In this thesis, a new method for protein structure determination – electron crystallography of multilayered protein crystals – has been successfully applied for the first time. The thesis is the experimental proof of the more theoretically oriented work initiated by Burmester, Dimmeler and Vossen. We have shown that the information about the protein structure can be extracted from diffraction data of multilayered protein crystals. In our work, a 3D dataset was recorded from the multilayered catalase crystals to 4Å resolution and the protein structure was successfully determined by Molecular Replacement. Although the new method has been developed now to the stage when immediate use is possible, collection of experimental diffraction data and their processing are still tedious tasks. To improve data quality and accelerate data collection, the following improvements are required. Firstly, it is necessary to develop a new crystallization method which will increase significantly the crystal concentration on the EM grid. Secondly, the radiation damage of the crystal can be decreased dramatically if high energy sources, for example 200 kV and 300 kV, are used instead of a 120 kV electron microscope as in our current study.

Furthermore, to get electron crystallography of multilayered crystals widely used, it must be faster in speed of data collection to make it more practical. It will require using modern technologies and all-round automation. For example, liquid helium (4 K) should be used instead of liquid nitrogen (77 K) to minimize beam damage of the sample. A new automated cryoholder which performs automatic tilt correction, should be coupled to a robotic system for 24 hours to provide unattended use of the microscope. In addition, a new generation of detectors, e.g. pixel detectors, with higher dynamic range will allow on-line evaluation of diffraction patterns at early stages of data collection.

All these developments will decrease the data collection time from weeks and even months to a few hours or days. For example, using a new cryoholder system and liquid helium cooling could allow to record 10-15 diffraction patterns from on catalase crystal. It means that only 50 crystals will be required to collect complete 3D dataset up to 3 Å.

Appendix A

Materials and Methods

A.1 Crystallization Experiments

A.1.1 Catalase Crystals

A suspension of bovine liver catalase crystals (Serva, catalog number 26900, 61312 U/mg) was centrifuged (14000 rpm, 5 min) and the pellet was dissolved in 20 % NaCl, 0.04 % NaN_3. The catalase volume solution (5-7 ml) was dialyzed (dialysis tubing MW cut-off 14000) against a 30 mM K^+-phosphate buffer pH 6.3 (volume 0.5 l) and stored at 4°C. Every 12 hours a quarter of the buffer (125 ml) was replaced by pure water. After a few days crystals appeared, as judged by the silkiness of the solution. The crystals have the symmetry of the primitive space group $P2_12_12_1$ with unit cell parameters a=69 Å, b=173 Å, c=206 Å, $\alpha=\beta=\gamma=90°$ (103). After 7 days the dialysis was stopped and the crystal suspension was stored at 4°C.

A.1.2 ClpB Crystallization

Microcrystals were grown by the sitting drop method by mixing an equal volume (2 µl) of proteins solution (8 mg/ml) in 50 mM Tris/HCl pH 7.5, 50 mM KCl, 5 mM $MgCl_2$, 2 mM EDTA, 2 mM DTE, 10 % (v/v) glycerine with reservoir solution consisting of 0.3-0.7 M magnesium formate. Crystals appeared within 5 hours.

Using the sitting drop geometry (Fig. A.1) as opposed to the large scale dialysis method or the hanging drop geometry has an advantage that no transfer steps of the crystals to the grid are required.
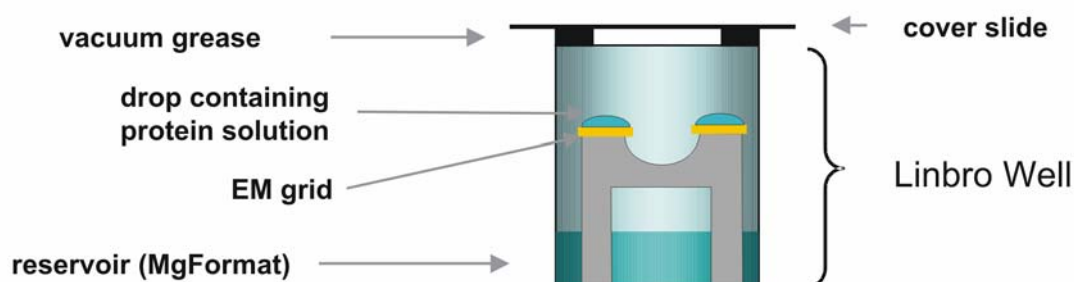

Figure A.1: ClpB crystallization. To avoid the problem of transferring the crystals onto the EM grid, crystallization is performed directly on the grid.

A.2 Grid Preparation

A.2.1 Amorphous Carbon Films

Amorphous carbon films were prepared by depositing 30-100 nm thick layers of carbon onto the copper supporting grids. The grids were first coated with a thin layer of a polymer called collodium. Then the grids were transferred into a vacuum chamber where carbon is evaporated by applying a voltage of about 150 V to a carbon thread spanned between two electrodes.

A.2.2 Making Grids for Cryo-microscopy

4-5 μL of catalase crystal suspension was pipetted on carbon film coated EM grid. In the case of ClpB, crystallization occurred directly on the grid and, therefore, the grid with protein suspension could be used for the blotting procedure immediately after opening the

cover slide (Figure A.1). The excess of the solution was blotted by a filter paper. After blotting the carbon film with the thin layer of crystal suspension was flash-cooled in a liquid ethane-propane (50/50 %) mixture.

A.3 Electron Microscopy (EFTEM) and Data Collection

All electron diffraction patterns were recorded using a LEO EFTEM 912 Ω at 120 keV energy. Specimen were held at a temperature of 101 K in the EM using an Oxford Cryo Transfer System CT 3500 Z. To record zero-loss energy diffraction patterns, the energy width of the filter was set to 25 eV. To protect the sample from unnecessary beam damage, the condenser aperture system (diameter 150 μm) was used to illuminate only a small part of the sample. The exposure was carried out under low electron dose condition with a dose of 2.5 e-/Å^2 for one pattern. For collection of diffraction patterns the camera length was 3600 mm, using an illumination aperture of 0.04 mrad. The crystals were sorted by size and diffraction data were recorded at different tilt angles from crystals larger than 2 μm in one dimension. Typically one tilt series consisted of 3-4 diffraction patterns.

A.4 Detection Media

All diffraction patterns were recorded on Fuji Imaging Plates. The plates are coated with a detection surface made of $BaFBr:Eu^{2+}$. When an electron hits this surface, it forms an electron-hole pair. This pair holds the image information until the plate is read in the scanning device.

The Imaging Plates were scanned at a pixel resolution of 25*25 μm^2 using the DITABIS Imaging Plate scanner. Laser light with a wavelength of 623 nm is used to recombine the an electron-hole pairs in the electron sensitive detection surface. The recombination leads

to the emission of 390 nm light which is detected using a photomultiplier. The detected light intensity corresponds to the grey value of the recorded image. The region of linearity is several orders of magnitude which allows the recording of 16 bit of data per pixel. The light detected by the photomultiplier is recorded in two independent images referred to as *high* and *low*. High light intensities are recorded in the *low* image, low intensities are recorded in the *high* image. This recording scheme allows the analysis of the complete dynamic range offered by the Imaging Plates.

A.5 Data Processing

The programs described in this chapter were used for processing electron diffraction images (Chapter 4). The overview of all steps is provided in the flowchart in figure A.2.
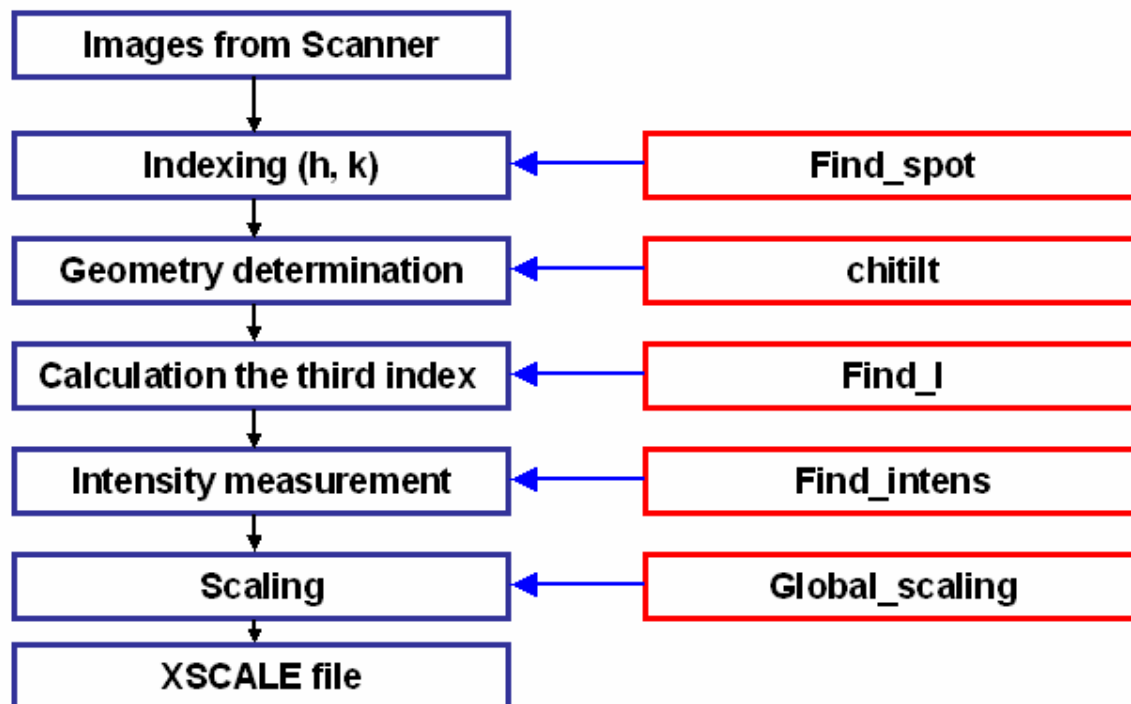

Figure A.2: Flowchart of diffraction data processing.

A.5.1 Spot Positions

To find diffraction spot positions, the program find_spot developed by Vossen was used with minor modifications (*21*). The user is prompted to enter the approximate spot positions of two Friedel pairs to calculate an origin of the pattern and the coordinates of the two lattice vectors. The result is a text file of the spots in the diffraction pattern and an image file with found spot positions (circles) superimposed on the diffraction pattern.

A.5.2 Distortion Correction

When position of diffraction spots are determined by find_spot, a correction of these position must be performed. The reason for this correction is a distortion of the diffraction patterns due to non-ideality of the DITABIS scanner. If an object is a circle (Fig.A.3), the scanned image of the object is an ellipse. It means that each point of the ellipse (B) must be
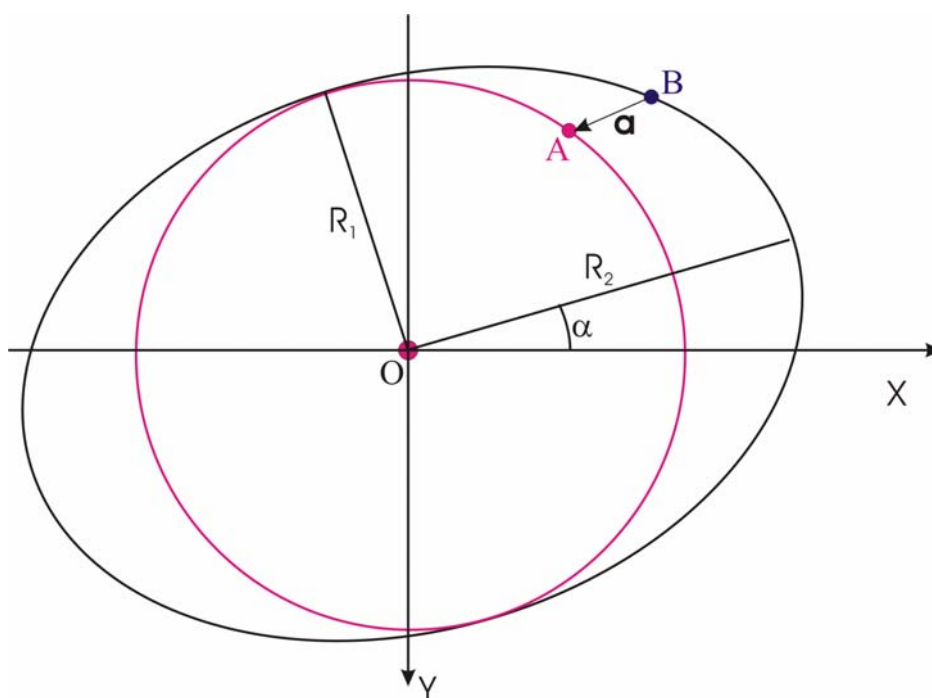

Figure A.3: Distortion correction of electron diffraction patterns.

shifted by a vector **a** to the point A to obtain an undistorted image, i.e. the circle.

The correction algorithm is realized in the program correction_x_direction (Kulik). The user is prompted to enter a name of the spot position file (usually, com.pos, find_spot output)[21], the angle between the axis x and the longest radius R_2 of the ellipse (α=-16°, Spider system of coordinates) and a ratio of R_1/R_2=0.934. The last two values are empirical and were determined from the elongation of the water ring on diffraction patterns. If different types of the Imaging Plates Scanner will be used, new values of α and R_1/R_2 must be calculated.

A.5.3 Crystal Orientation

The fit program chitilt (*23;24*) was used to determine the orientation of the crystal at high tilt geometry (>10°). The seven parameters needed for a complete orientation of the crystal position (see figure 4.6) are calculated using an unweighted least-squares fit algorithm. An iterative approach is used to minimize the sum of the mean errors χ^2.

To determine the tilt geometry of low tilted crystals (<10°), the Laue zone analysis was carried out. For these purposes the following steps have to be performed:

1) Using a text file (output find_intens), the user should create a SPIDER file, simply by assigning each diffraction spot intensity to a pixel with different grey value in the SPIDER image.

2) Since the spacing between the diffraction spots for bovine catalase crystals have to correspond to **a***=69 Å and **b***=173 Å, the SPIDER image should be interpolated in one direction by the factor 173/69. Now the Laue zones should be visible and correspond to the original geometry in the diffraction pattern.

3) The next step is to rotate the SPIDER image by different angles and project the rotated image onto the one-dimensional grey value distribution. The projection has

---

[21] N.B. Do not forget to delete the first two lines and the line which contains information about the 0,0 reflection. After correction insert these lines again. For more information see the program code.

to show intensity variations (peaks) along the axis due to presence of the Laue zones. These peaks have to be maximal values when the rotated SPIDER image is parallel to the Laue zone. The projection with the strongest intensity peaks corresponds to the direction of the Laue zones (angle Φ).

4) The distances between the peak maxima correspond to the distances between the Laue zones. Therefore, the angle Θ can be determined by considering the diffraction geometry (intersection of the Ewald sphere with the first, second and so on lattice layer).

A.5.4 The Third Miller Index l

The program find_l calculates the nearest Miller index according to the crystal tilt geometry. The output text file contains positions of spots (in pixels) and three Miller indexes.

A.5.5 Intensity Recording

The program find_intens is used to determine the intensities of the diffraction spots. This algorithm does not calculate a background corrected image but merely subtracts the background of a given spot from the integrated intensity.

A.5.6 Scaling

To compare the intensities of diffraction patterns, the individual spot intensities should be normalized. This is done by calculating the sum of all diffraction spot intensities and dividing each intensity by that value. The program global_scaling is the implementation of this algorithm. The output is a file in the XSCALE format.

A.6 Molecular Replacement

The flowchart in figure A.4 presents an overview of the steps necessary to solve the phase problem by Molecular replacement and to analyze missing cone and data incompleteness effects on MR, NCS and density reconstruction.

A.6.1 Datasets with Different Completeness

To create datasets with different completeness, the programs isotropic_data and missing_cone were used. The program isotropic_data omits a given percentage of diffraction spots in a random manner. The program missing_cone removes only those reflections which belong to a cone oriented in the Z-direction with the vertex of the cone placed at (0,0,0) and a given cone angle.
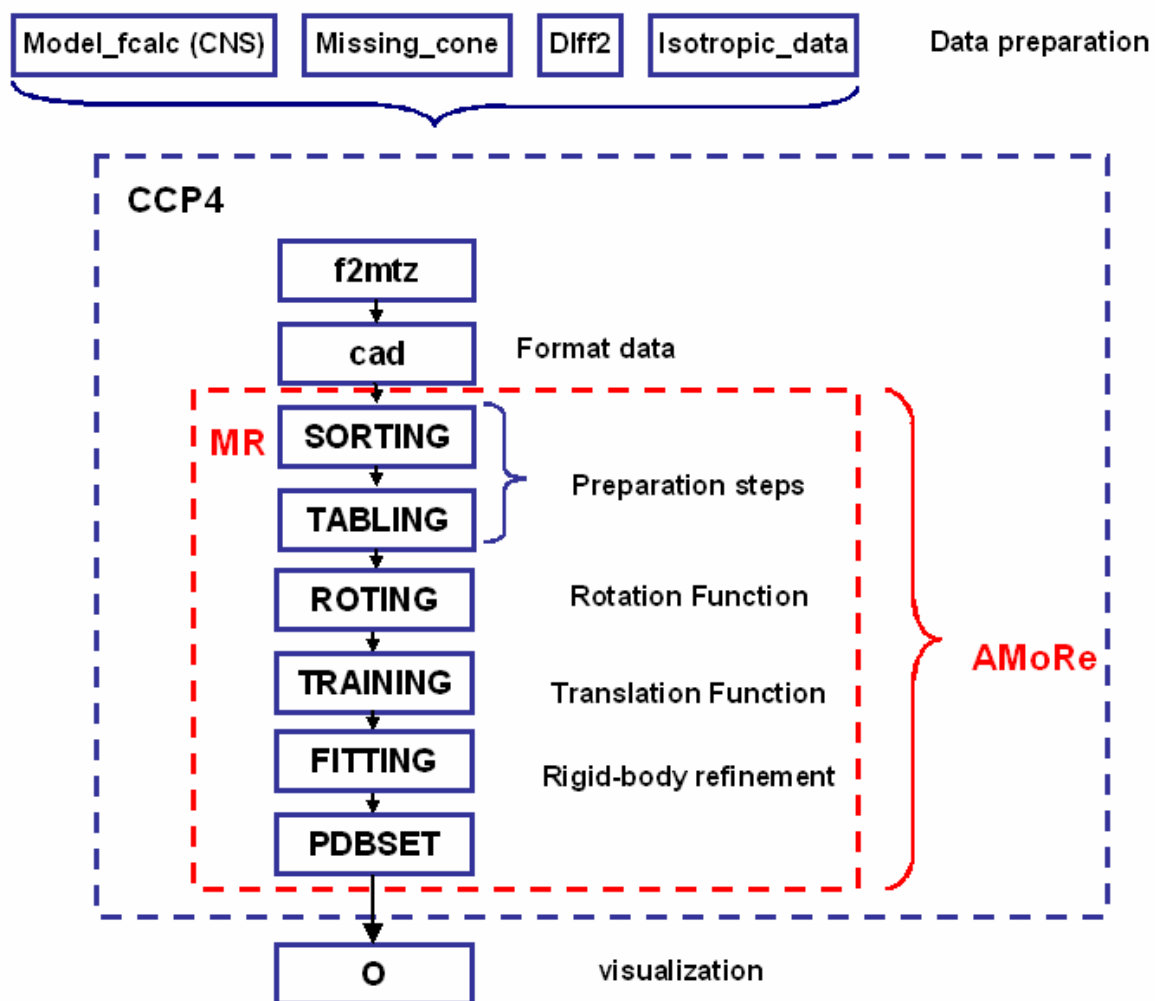
Figure A.4: Flowchart of the Molecular Replacement steps using the CCP4 program package.

Bibliography

1. Benedicto Spinoza (1677) *Ethics, Pt. IV, Prop. LXVII.*

2. Schroedinger E. (1944) *What is Life; The Physical Aspect of the Living Cell* Cambridge University Press, Cambridge.

3. Radzicka, A. and Wolfenden, R. (1995) *Science 267*, 90-93.

4. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di, F., V, Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C.,

Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., and Nodell, M. (2001) *Science 291*, 1304-1351.

5. Sali, A., Glaeser, R., Earnest, T., and Baumeister, W. (2003) *Nature 422*, 216-225.

6. Kuntz, I. D. (1992) *Science 257*, 1078-1082.

7. Somorjai, R. L., Nikulin, A. E., Pizzi, N., Jackson, D., Scarth, G., Dolenko, B., Gordon, H., Russell, P., Lean, C. L., Delbridge, L., and . (1995) *Magn Reson. Med. 33*, 257-263.

8. Moller, H., Serttas, N., Paulsen, H., Burchell, J. M., and Taylor-Papadimitriou, J. (2002) *Eur. J. Biochem. 269*, 1444-1455.

9. Austin, R., Buhks, E., Chance, B., D.De Vault, Dulton P.L., Frayenfelder H., and Gol'danskii V.I. (1987) *Protein Structure: Molecular and Electronic Reactivity* Springer-Verlag, New York.

10. Schlichting, I. (2000) *Acc. Chem. Res. 33*, 532-538.

11. Perman, B., Anderson, S., Schmidt, M., and Moffat, K. (2000) *Cell Mol. Biol. (Noisy. -le-grand) 46*, 895-913.

12. Moffat, K. (1998) *Acta Cryst. 54()*, 833-841.

13. Schertler, G. F., Villa, C., and Henderson, R. (1993) *Nature 362*, 770-772.

14. Kuhlbrandt, W., Wang, D. N., and Fujiyoshi, Y. (1994) *Nature 367*, 614-621.

15. Auer, M., Scarborough, G. A., and Kuhlbrandt, W. (1998) *Nature 392*, 840-843.

16. Koning, R. I., Keegstra, W., Oostergetel, G. T., Schuurman-Wolters, G., Robillard, G. T., and Brisson, A. (1999) *J. Mol. Biol. 287*, 845-851.

17. Shi, D., Lewis, M. R., Young, H. S., and Stokes, D. L. (1998) *J. Mol. Biol. 284*, 1547-1564.

18. Stokes, D. L. and Green, N. M. (1990) *J. Mol. Biol. 213*, 529-538.

19. Brink, J., Chiu, W., and Dougherty, M. (1992) *Ultramicroscopy 46*, 229-240.

20. Grant, R. A., Schmid, M. F., and Chiu, W. (1991) *J. Mol. Biol. 217*, 551-562.

21. Vossen, O. Effects of dynamical scattering on electron Crystallography of multi-layered crystalline biological samples. 2002. University of Heidelberg.
Ref Type: Thesis/Dissertation

22. Dimmeler, E., Vossen, O., and Schroder, R. R. (2000) *Journal of Applied Crystallography 33*, 1102-1112.

23. Dimmeler, E. and Schroder, R. R. (2000) *Journal of Applied Crystallography 33*, 1088-1101.

24. Dimmeler, E. Bestimmung der Steuamplituden bei der electronenkristallographischen Strukturuntersuchung von 3D-Protein-Mikrokristallen. 1998.
Ref Type: Thesis/Dissertation

25. Reimer, L. (1989) *Transmission Electron Microscopy* Springer-Verlag.

26. Blundell, T. L. and Johnson, L. N. (1976) *Protein Crystallography*.

27. Henderson, R. (1995) *Q. Rev. Biophys 28*, 171-193.

28. Frank, J. (1998) *J. Struct. Biol. 124*, 142-150.

29. Glaeser, R. M. and Ceska, T. A. (1989) *Acta Cryst. 45*, 620-628.

30. Penczek, P., Radermacher, M., and Frank, J. (1992) *Ultramicroscopy 40*, 33-53.

31. Unger, V. M. (2001) *Curr. Opin. Struct. Biol. 11*, 548-554.

32. Grigorieff, N., Ceska, T. A., Downing, K. H., Baldwin, J. M., and Henderson, R. (1996) *J. Mol. Biol. 259*, 393-421.

33. Nogales, E., Wolf, S. G., and Downing, K. H. (1998) *Nature 391*, 199-203.

34. Walz, T., Jamieson, S. J., Bowers, C. M., Bullough, P. A., and Hunter, C. N. (1998) *J. Mol. Biol. 282*, 833-845.

35. Jap, B. K. and Li, H. (1995) *J. Mol. Biol. 251*, 413-420.

36. Schmidt-Krey, I., Murata, K., Hirai, T., Mitsuoka, K., Cheng, Y., Morgenstern, R., Fujiyoshi, Y., and Hebert, H. (1999) *J. Mol. Biol. 288*, 243-253.

37. Auer, M. (2000) *Journal of Molecular Medicine-Jmm 78*, 191-202.

38. Sakai, H. and Tsukihara, T. (1998) *J. Biochem. (Tokyo) 124*, 1051-1059.

39. Pautsch, A. and Schulz, G. E. (1998) *Nat. Struct. Biol. 5*, 1013-1017.

40. Snijder, H. J., Ubarretxena-Belandia, I., Blaauw, M., Kalk, K. H., Verheij, H. M., Egmond, M. R., Dekker, N., and Dijkstra, B. W. (1999) *Nature 401*, 717-721.

41. Buchanan, S. K., Smith, B. S., Venkatramani, L., Xia, D., Esser, L., Palnitkar, M., Chakraborty, R., van der, H. D., and Deisenhofer, J. (1999) *Nat. Struct. Biol. 6*, 56-63.

42. Weber, P. C. (1991) *Adv. Protein Chem. 41*, 1-36.

43. Drenth, J. and Haas, C. (1992) *Journal of Crystal Growth 122*, 107-109.

44. McPherson, A., Malkin, A. J., and Kuznetsov, Y. G. (1995) *Structure 3*, 759-768.

45. Boistelle, R. and Astier, J. P. (1988) *Journal of Crystal Growth 90*, 14-30.

46. Sato, K., Fukuba, Y., Mitsuda, T., Hirai, K., and Moriya, K. (1992) *Journal of Crystal Growth 122*, 87-94.

47. POLSON, A., POTGIETER, G. M., LARGIER, J. F., MEARS, G. E., and JOUBERT, F. J. (1964) *Biochim. Biophys. Acta 82*, 463-475.

48. McPherson, A., Jr. (1976) *J. Biol. Chem. 251*, 6300-6303.

49. Chirgadze, D. Protein crystallisation in action. http://www-cryst.bioc.cam.ac.uk/~dima/xtal-in-action/index.html . 2001.
Ref Type: Electronic Citation

50. Mosser, G. (2001) *Micron 32*, 517-540.

51. Miyazawa, A., Fujiyoshi, Y., and Unwin, N. (2003) *Nature 424*, 949-955.

52. Auer, M., Scarborough, G. A., and Kuhlbrandt, W. (1999) *J. Mol. Biol. 287*, 961-968.

53. Darwin, C. G. (1922) *Philosophical Magazine 43*, 800-829.

54. SCHNUCHEL, G. (1954) *Hoppe Seylers. Z. Physiol Chem. 298*, 241-244.

55. Dubochet, J., Adrian, M., Chang, J. J., Homo, J. C., Lepault, J., McDowall, A. W., and Schultz, P. (1988) *Q. Rev. Biophys. 21*, 129-228.

56. Schroder, R. R. (1992) *Journal of Microscopy 166*, 389-400.

57. Angert, I., Majorovits, E., and Schroder, R. R. (2000) *Ultramicroscopy 81*, 203-222.

58. Ko, T. P., Day, J., Malkin, A. J., and McPherson, A. (1999) *Acta Crystallogr. D Biol. Crystallogr. 55 (Pt 8)*, 1383-1394.

59. Collaborative Computational Project, N. 4. (1994) *Acta Cryst. D50*, 760-763.

60. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) *Acta Crystallogr. D Biol. Crystallogr. 54 (Pt 5)*, 905-921.

61. Engel, A. and Massalski, A. (1984) *Ultramicroscopy 13*, 71-83.

62. Prasad, B. V., Degn, L. L., Jeng, T. W., and Chiu, W. (1990) *Ultramicroscopy 33*, 281-285.

63. Weiss, M. S. (2001) *Journal Of Applied Crystallography 34*, 130-135.

64. Weiss, M. S., Metzner, H. J., and Hilgenfeld, R. (1998) *FEBS Letters 423*, 291-296.

65. Diederichs, K. and Karplus, P. A. (1997) *Nat. Struct. Biol. 4*, 269-275.

66. Duncan E.McRee (1993) *Pracrical protein crystallography* Academic Press, Inc..

67. Rossmann, M. G. and Blow, D. M. (1962) *Acta Crystallographica 15*, 24-&.

68. Bricogne, G. (1976) *Acta Cryst. 32*, 832-847.

69. Crowther, R. A. (1972) in *The Molecular Replacement Method* pp 173-178, Gordon&Breach, New York.

70. Navaza, J. (1994) *Acta Cryst. 50*, 157-163.

71. Navaza, J. (1987) *Acta Cryst. 43*, 645-653.

72. Navaza, J. (1990) *Acta Cryst. 46*, 619-620.

73. Castellano, E. E., Oliva, G., and Navaza, J. (1992) *Journal of Applied Crystallography 25*, 281-284.

74. Rossmann, M. G. and Blow, D. M. (1967) *Acta Cryst. 23*, 548-&.

75. Read, R. J. and Schierbeek, A. J. (1988) *Journal of Applied Crystallography 21*, 490-495.

76. Zhang, X.-J. and Matthews, B. W. (1994) *Acta Cryst. D50*, 675-686.

77. Brunger, A. T. (1990) *Acta Cryst. 46*, 46-57.

78. Navaza, J. and Vernoslova, E. (1995) *Acta Cryst. 51*, 445-449.

79. Deisseroth, A. and Dounce, A. L. (1970) *Physiol Rev. 50*, 319-375.

80. Jones, P. and Suggett, A. (1968) *Biochem. J. 108*, 833-838.

81. Matthews, B. W. (1968) *J. Mol. Biol. 33*, 491-497.

82. Ceska, T. A. (1994) *Journal of Applied Crystallography 27*, 200-201.

83. Bailey, S. (1994) *Acta Crystallographica Section D-Biological Crystallography 50*, 760-763.

84. Huber, R. and Schneider, M. (1985) *Journal Of Applied Crystallography 18*, 165-169.

85. Jones, T. A., Zou, J. Y., Cowan, S. W., and Kjelgaard, M. (1991) *Acta Cryst. A47*, 110-119.

86. Burmester, C. and Schroder, R. R. (1997) *Scanning Microscopy 11*, 323-334.

87. Vossen, O. Effects of dynamical scattering on electron Crystallography of multi-layered crystalline biological samples. 2003. 3-7-2002.
Ref Type: Thesis/Dissertation

88. Burmester, C. Neue Methoden zur quantifizierung von Electronendiffraktogrammen und ihre Anwendung auf die Phasenbestimmung bei Proteinkristallen ьber isomorphen Scheratomersatz. 1996.
Ref Type: Thesis/Dissertation

89. McPherson, A., Jr. and Rich, A. (1973) *Arch. Biochem. Biophys. 157*, 23-27.

90. Ceska, T. A. and Henderson, R. (1990) *J. Mol. Biol. 213*, 539-560.

91. Sali, A. (1995) *Curr. Opin. Biotechnol. 6*, 437-451.

92. Lee, S., Sowa, M. E., Watanabe, Y. H., Sigler, P. B., Chiu, W., Yoshida, M., and Tsai, F. T. (2003) *Cell 115*, 229-240.

93. Schlee, S., Groemping, Y., Herde, P., Seidel, R., and Reinstein, J. (2001) *J. Mol. Biol. 306*, 889-899.

94. Carragher, B., Kisseberth, N., Kriegman, D., Milligan, R. A., Potter, C. S., Pulokas, J., and Reilein, A. (2000) *J. Struct. Biol. 132*, 33-45.

95. Potter, C. S., Chu, H., Frey, B., Green, C., Kisseberth, N., Madden, T. J., Miller, K. L., Nahrstedt, K., Pulokas, J., Reilein, A., Tcheng, D., Weber, D., and Carragher, B. (1999) *Ultramicroscopy 77*, 153-161.

96. Perrakis, A., Cipriani, F., Castagna, J. C., Claustre, L., Burghammer, M., Riekel, C., and Cusack, S. (1999) *Acta Crystallographica Section D-Biological Crystallography 55*, 1765-1770.

97. J.Struct.Biol. 145[1-2], 1-180. 2004.
Ref Type: Journal (Full)

98. Faruqi, A. R., Cattermole, D. M., Henderson, R., Mikulec, B., and Raeburn, C. (2003) *Ultramicroscopy 94*, 263-276.

99. Zemlin, F. (1992) *Ultramicroscopy 46*, 25-32.

100. Stahlberg, H., Fotiadis, D., Scheuring, S., Remigy, H., Braun, T., Mitsuoka, K., Fujiyoshi, Y., and Engel, A. (2001) *FEBS Letters 504*, 166-172.

101. Fujiyoshi, Y., Mizusaki, T., Morikawa, K., Yamagishi, H., Aoki, Y., Kihara, H., and Harada, Y. (1991) *Ultramicroscopy 38*, 241-251.

102. Koster, A. J., Grimm, R., Typke, D., Hegerl, R., Stoschek, A., Walz, J., and Baumeister, W. (1997) *J. Struct. Biol. 120*, 276-308.

103. Unwin, P. N. (1975) *J. Mol. Biol. 98*, 235-242.

Publications and Poster Presentations:

Kulik V., Weyand M., Seidel R., Schneider T.R., F. Dunn M.F., Schlichting I. (2001) : Structural Basis for the Inhibited Inter-subunit Communication in a Tryptophan Synthase Mutant from *Salmonella Typhimurium, Proc. 5th International School on the Crystallisation of Biological Macromolecules*, Como, Italy

Kulik V, Weyand M, Seidel R, Niks D, Arac D, Dunn M.F, Schlichting I. (2002) : On the Role of alphaThr183 in the Allosteric Regulation and Catalytic Mechanism of Tryptophan Synthase, *J. Mol. Biol.* **324(4)** pp.677-690

Kulik V., Schlichting I.: Structural Comparison of a BX1 Enzyme and the Alpha-subunit of a Tryptophan Synthase, in preparation

Kulik V., Vossen O., Majorovits E., Schlichting I., Schröder R.R. (2003): Structure of Bovine Liver Catalase Solved from Electron Diffraction Data of Multilayered Crystals, *Proc. Microscopy Conference*, Dresden, Germany

Kulik V., Schröder R.R. : Structure of Bovine Liver Catalase Solved by Electron Diffraction on Multilayered Crystals and Molecular Replacement, in preparation

Acknowledgment