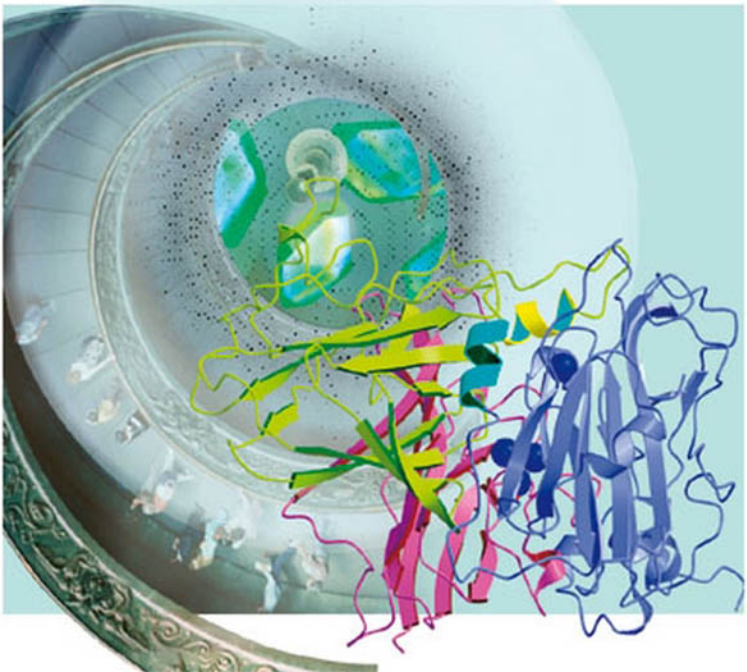
Albrecht Messerschmidt

WILEY-VCH

X-Ray Crystallography of Biomacromolecules

A Practical Guide



Albrecht Messerschmidt

X-Ray Crystallography of Biomacromolecules

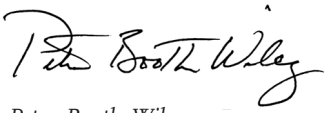
1807–2007 Knowledge for Generations

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

William J. Pesce
President and Chief Executive Officer



Peter Booth Wiley
Chairman of the Board

Albrecht Messerschmidt

X-Ray Crystallography of Biomacromolecules

A Practical Guide

WILEY-VCH Verlag GmbH & Co. KGaA

The Author
Dr. Albrecht Messerschmidt
Max-Planck-Institue of Biochemistry
Department of Proteomics and Signal Transduction
Research Group: Structural Proteomics
Am Klopferspitz 18
82152 Martinsried
Germany
Phone: +49 89 8578 2669
Fax: +49 89 8578 2219
E-Mail: messersc@biochem.mpg.de

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Typesetting K+V Fotosatz GmbH, Beerfelden
Printing Strauss GmbH, Darmstadt
Bookbinding Litges & Dopf Buchbinderei GmbH, Heppenheim
Cover Design Adam-Design, Weinheim

Printed in the Federal Republic of Germany
Printed on acid-free paper

ISBN 978-3-527-31396-9

Dedicated to my scientific teachers
Will Kleber
Katharina Boll-Dornberger
Ernst Höhne
Robert Huber

and to my wife
Beate

Contents

Preface

Knowledge of the atomic structure of biomacromolecules such as proteins, nucleic acids and carbohydrates is indispensable for an understanding their biological function. The sequencing of whole genomes from bacteria to man has provided the possibility to clone and express their gene products, which in most cases are proteins. The new field of Structural Genomics, or its synonym Structural Proteomics, aims at determining the three-dimensional structures of the whole set of gene products – the so-called targets – or of a biological or medical important subset of targets. X-ray crystallography is the major technique to determine the atomic structure of biomacromolecules. This book provides the theoretical basis and covers the experimental techniques of X-ray crystal structure determination of such molecules. It documents the state-of-the-art of this method, including practical case studies. It will be of excellent use for students and researchers active in this field, and will also serve as source of information for readers from other related research areas.

This book is not thought of as a simple practical guide, but rather to supply the theoretical basis necessary to understand the underlying physics and mathematics of the diffraction of X-rays and the related crystal structure determination of biomacromolecules. The intention is to form a platform to run the many automated procedures in X-ray crystallography of biomacromolecules, and to be aware of the great science and technique which is contained in the technical equipment, prescriptions, and expert computer software systems.

I would like to thank my former director Robert Huber for giving me the opportunity to work in his Department of Structural Research at the Max-Planck-Institute of Biochemistry in Martinsried, and to become familiar with the ever-fascinating methods of X-ray crystallography of biomacromolecules. I also wish to express my gratitude to Peter Kroneck for many successful common projects, and the possibility to act as external teacher at the Biological Faculty of the University of Constance. I am cordially indebted to my wife Beate who sacrificed plenty of her precious time to prepare a substantial part of the drawings.

Martinsried, September 2006 *Albrecht Messerschmidt*

Part I
Principles and Methods

1
Introduction

1.1
Crystals and Symmetry

Who has not been fascinated by the regular shape of single crystals of minerals, gemstones, other inorganic compounds and organic substances? Yet, most biological macromolecules can also be crystallized. A characteristic of the so-called morphology of crystals is a set of flat faces, forming a closed body. Figure 1.1 shows a regularly shaped quartz crystal, but the shape may also be skewed, as depicted in Figure 1.2, for lodestone (magnetite, Fe_3O_4).

It was first shown by Nicolaus Steno (in 1669) that the angles between the faces are constant, independently of the regularity of a given crystal morphology. The analysis of crystal morphologies led to the formulation of a complete set of 32 symmetry classes – also called "point groups" – which all crystal morphologies obey. Possible symmetry elements are 1-, 2-, 3-, 4-, and 6-fold rotations, mirror plane *m*, inversion center and a combination of rotation axis with inversion center (inversion axis). As explained later, crystals of biological macromolecules can contain rotation symmetries only, thereby reducing the possible point groups to the 11 enantiomeric point groups: 1, 2, 3, 4, 6, 222, 32, 422, 622, 23, and 432. A graphical representation of the symmetries and of their general morphological crystal form is displayed in Figure 1.3.

The morphology of a crystal tells us much about its symmetry, but little about its internal structure. Before the discovery of X-ray diffraction of crystals by von Laue, Knipping and Friedrich in 1912, it had been proven that crystals are built up from atoms or molecules arranged in a three-dimensionally periodic

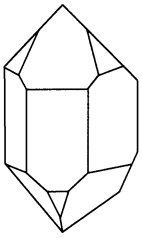


Fig. 1.1 Regularly shaped quartz crystal.

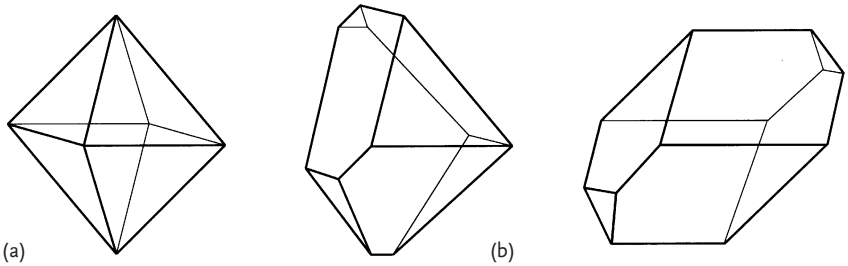


Fig. 1.2 Different forms of the octahedron of lodestone. (a) Regular shape; (b) skewed shape caused by parallel shift of faces.

manner by translational symmetry. The crystal is formed by a three-dimensional stack of unit cells which is called the "crystal lattice" (Fig. 1.4). The unit cell is built up from three noncolinear vectors **a**, **b**, and **c**. In the general case, these vectors have unequal magnitudes and their mutual angles deviate from 90°. The arrangement of the molecule(s) in the unit cell may be asymmetrical, but very often it is symmetrical. This is illustrated in Figure 1.5 in two-dimensional lattices for rotational symmetries.

It follows from the combination of the lattice properties with rotational operations that in crystals only 1-, 2-, 3-, 4-, and 6-fold axes are allowed, and they can only occur among each other in a few certain combinations of angles as other angle orientations would violate the lattice properties. The number of all possible combinations reveals the 32 point groups, and delivers the deviation of the 32 point groups on the basis of the symmetry theory.

Adding an inversion center to the point group symmetry leads to the 11 Laue groups. These are of importance for the symmetry of X-ray diffraction patterns. Their symbols are: 1, 2/*m*, 2/*mmm*, 3, 3*m*, 4/*m*, 4/*mmm*, 6/*m*, 6/*mmm*, *m*3, and *m*3*m*. Proteins and nucleic acids are chiral molecules and can, therefore, crystallize only in the 11 enantiomorphic point groups, as mentioned above.

The combination of point group symmetries with lattices leads to seven crystal systems, triclinic, monoclinic, orthorhombic, trigonal, tetragonal, hexagonal and cubic, the metric relationships of which are provided in Figure 1.6, with 14 different Bravais-lattice types which can be primitive, face-centered, all-face-centered, and body-centered (see Fig. 1.7). It is however possible to describe each translation lattice as a primitive lattice. Furthermore, different primitive unit cells can be chosen. Both situations are illustrated in Figure 1.8, where a two-dimensional (2D) face-centered tetragonal lattice is presented. The face-centered unit cell is assigned in the middle of the figure, and a primitive cell obeying the tetragonal symmetry has been marked by dashed lines. Three further putative primitive cells have been drawn in and numbered. Among these possible primitive unit cells or bases a so-called "reduced basis" **a**, **b**, **c** is important for the automated unit cell and space group determination of crystals from X-ray diffraction data. Such a basis is right-handed and the components of the metric tensor **G**

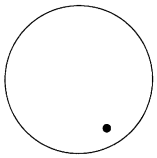
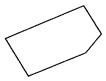
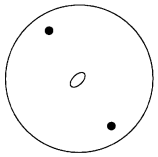
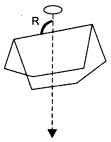
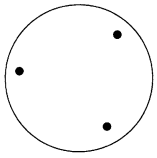
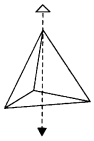
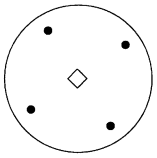
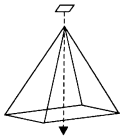
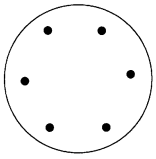
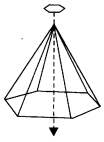
	symmetry framework	stereographic projection of the symmetry framework and of the general form	general form
C_1 1 triclinic-pedial pedion			
C_2 2 monoclinic-sphenoidal sphenoid			
C_3 3 trigonal-pyramidal trigonal pyramid			
C_4 4 tetragonal-pyramidal tetragonal pyramid			
C_6 6 hexagonal-pyramidal hexagonal pyramid			

Fig. 1.3 Graphical representation of the 11 enantiomorphic point groups.

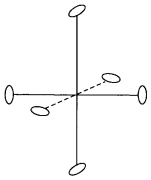
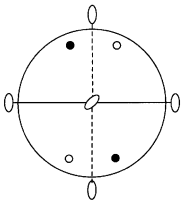
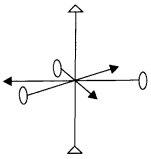
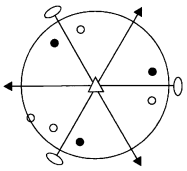
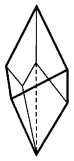
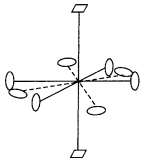
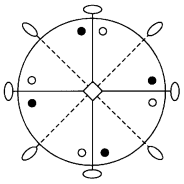
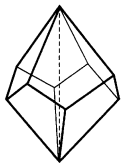
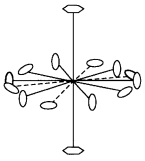
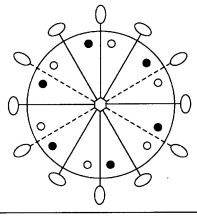
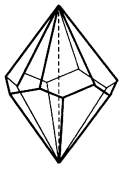
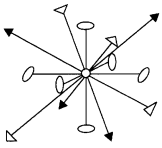
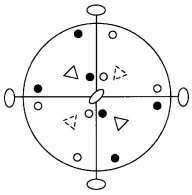
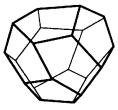
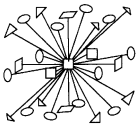
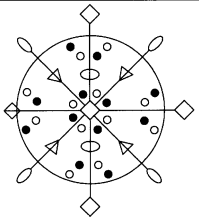
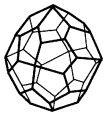
D_2 222 orthorhombic-disphenoidal disphenoid			
D_3 32 trigonal-traphezohedral trigonal-traphezohedron			
D_4 422 tetragonal-trapezohedral tetragonal-trapezohedron			
D_6 622 hexagonal-trapezohedral hexagonal-trapezohedron			
T 23 tetartoidal tetartoid			
O 432 gyroidal gyroid			

Fig. 1.3 (continued)

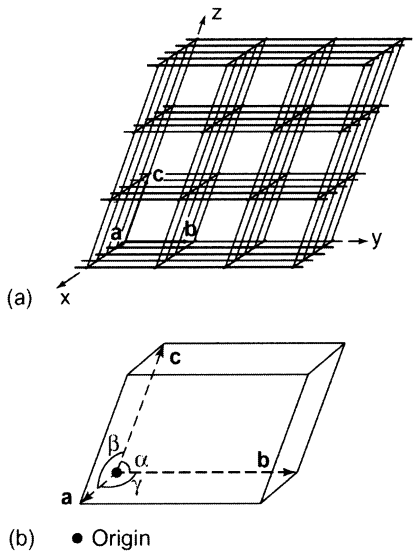


Fig. 1.4 (a) A crystal lattice; (b) a unit cell.

$$\begin{matrix} \mathbf{a}\cdot\mathbf{a} & \mathbf{b}\cdot\mathbf{b} & \mathbf{c}\cdot\mathbf{c} \\ \mathbf{b}\cdot\mathbf{c} & \mathbf{c}\cdot\mathbf{a} & \mathbf{a}\cdot\mathbf{b} \end{matrix} \tag{1.1}$$

satisfy special conditions listed in International Tables for Crystallography, Volume A, page 750 (Hahn, 2005).

Furthermore, additional symmetry elements are generated having translational components such as screw axes or glide mirror planes. There exist 230 space groups, of which 65 are enantiomorphic (for chiral molecules such as proteins); these are listed in Figure 1.9. As an example for such an additional symmetry element, the action of 3_1- and 3_2-screw axes is demonstrated in Figure 1.10. For a 3_1-axis, the object is rotated by 120° anti-clockwise and shifted by one-third of the translation parallel to the direction of the axis. This is repeated twice, and the rotational start position is reached but shifted by one translational unit, thus generating a right-handed screw axis. For a 3_2-axis, the object is again rotated by 120° anti-clockwise but shifted by two-thirds of the translation parallel to the direction of the axis. This is repeated twice and the rotational start position is reached but shifted by two translational units. The missing objects are obtained by applying the translation symmetry. The result is a left-handed screw axis. Figure 1.11 shows the graphical representation for the space group $P2_12_12_1$ as listed in the International Tables for Crystallography (Hahn, 2005). The asymmetric unit is one-fourth of the unit cell, and can contain one or several molecules. Multimeric molecules may have their own symmetries which are called noncrystallographic symmetries. Here, axes which are 5-fold, 7-fold, etc., are also allowed.

It is useful to describe the relationship between a crystal face and its counterpart in the crystal lattice. Figure 1.12a shows a crystal face in a general position inter-

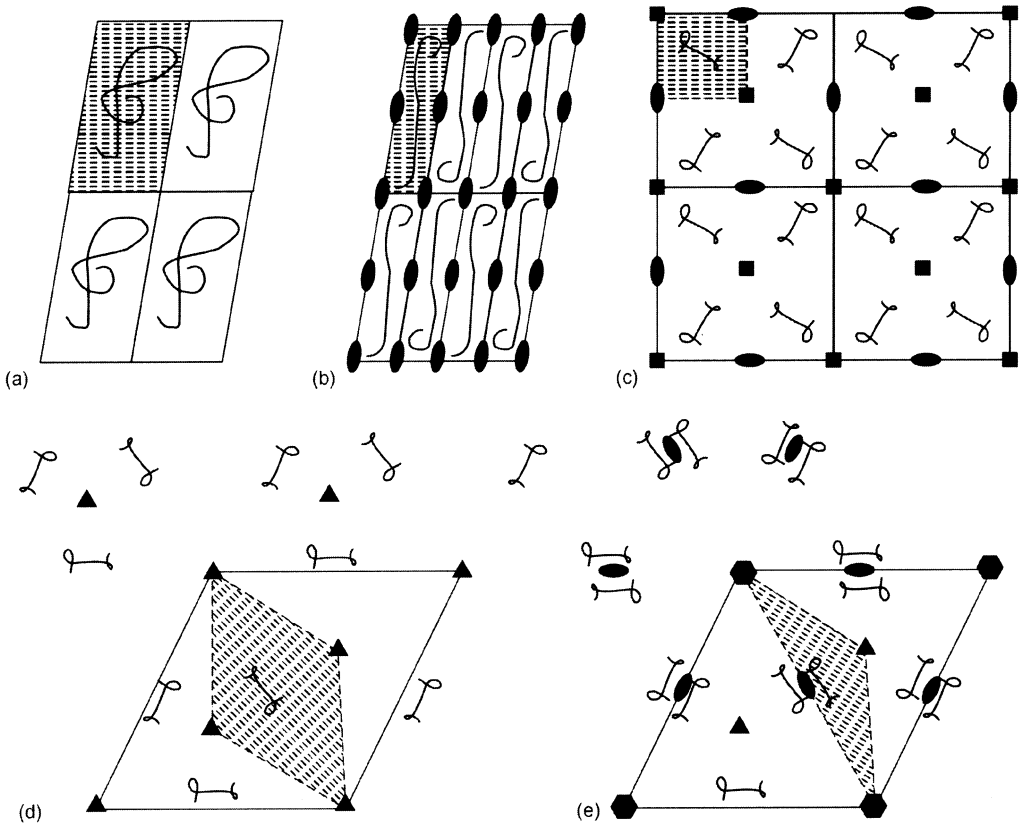


Fig. 1.5 Rotational symmetry elements in two-dimensional lattices: (a) 1, (b) 2, (c) 4, (d) 3, and (e) 62. The asymmetric unit is hatched.

Name	Possible Bravais Lattices	Axes of symmetry	Lattice	
Triclinic	*P*	No axes of symmetry	$a \neq b \neq c$	$\alpha \neq \beta \neq \gamma$
Monoclinic	*P, C*	1 dyad axis (parallel to b)	$a \neq b \neq c$	$\alpha = \gamma = 90° \neq \beta$
Orthorhombic	*P, C, I, F*	3 dyad axes mutually orthogonal	$a \neq b \neq c$	$\alpha = \beta = \gamma = 90°$
Tetragonal	*P, I*	1 tetrad axis (parallel to c)	$a = b \neq c$	$\alpha = \beta = \gamma = 90°$
Trigonal	*P* (or *R*)	1triad axis (parallel to c)	$a = b \neq c$	$\alpha = \beta = 90°, \gamma = 120°$
			$a = b = c$	$\alpha = \beta = \gamma < 120°, \neq 90)$
Hexagonal	*P*	1hexad axis (parallel to c)	$a = b \neq c$	$\alpha = \beta = 90°, \gamma = 120°$
Cubic	*P, I, F*	4 triad axes (along the diagonals of the cube)	$a = b = c$	$\alpha = \beta = \gamma = 90°$

Fig. 1.6 The seven crystal systems.

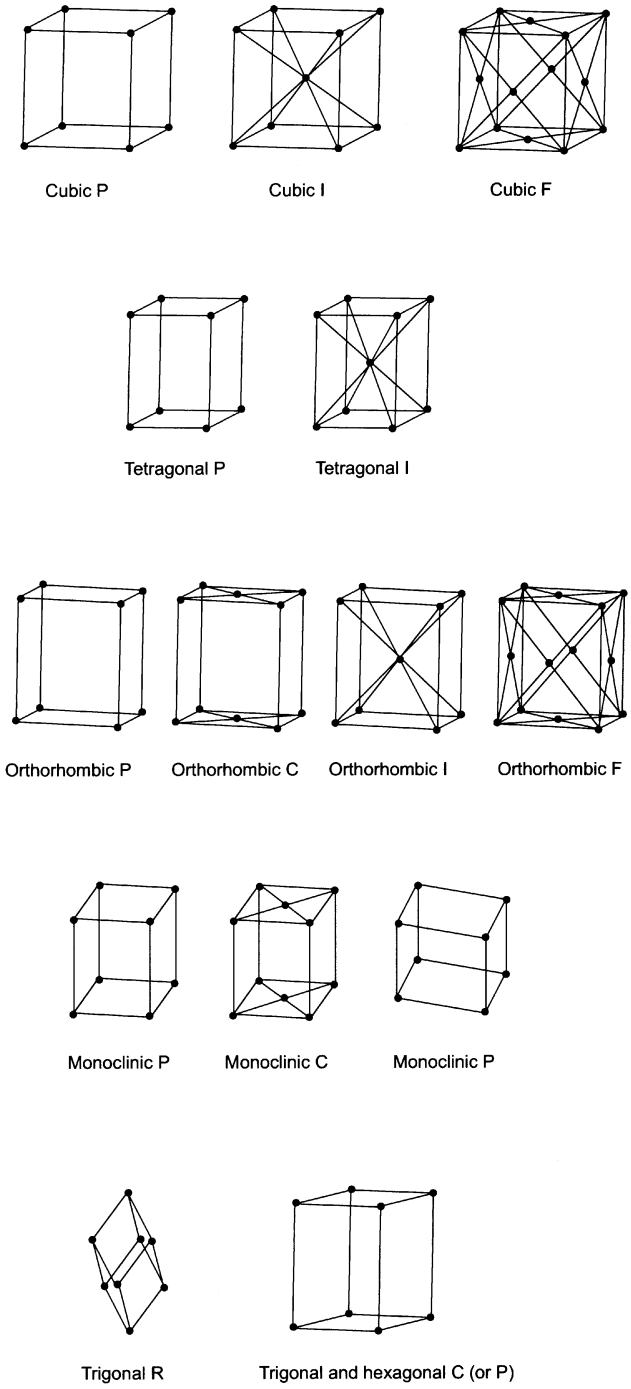


Fig. 1.7 The 14 Bravais lattices.

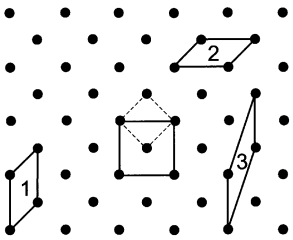


Fig. 1.8 Choice of different primitive unit cells.

secting the underlying coordinate system at distances OA, OB, OC on the a-, b-, and c-axes, respectively. Figures 1.12 b and c depict the relevant crystal lattice, but only in two dimensions for the sake of clarity. The counterpart of crystal faces are the lattice planes. In Figure 1.12 b the lattice planes have axis intercepts, which are in a ratio of 2*a* to 1*b*. In Figure 1.12 c the ratio is 2*a* to 3*b*. In general, we have *ma*, *nb*, *pc* with the rational numbers *m*, *n*, and *p*. In crystallography, it is not the axis intercepts but rather their reciprocal values which are used to characterize the position of a crystal face or lattice plane according to Eq. (1.2).

$$h : k : l = \frac{1}{m} : \frac{1}{n} : \frac{1}{p} \ . \tag{1.2}$$

The triple of numbers *h*, *k*, *l* is transformed in such a way that the numbers become integers and relatively prime. The *hkl* are called *Miller indices,* and can be applied to either crystal faces or lattice planes. The lattice planes are a stack of equidistant parallel planes with a lattice plane distance *d(hkl)*. The larger the *Miller indices* of a lattice plane, the smaller is *d(hkl)*.

Crystal system	Class	Point group symbols
Triclinic	1	*P*1
Monoclinic	2	*P*2, $P2_1$, *C*2
Orthorhombic	222	*C*222, *P*222, $P2_12_12_1$, $P2_12_12$, $P222_1$, $C22_1$, *F*222, *I*222, $I2_12_12_1$
Tetragonal	4	*P*4, $P4_1$, $P4_2$,$P4_3$, *I*4, $I4_1$
	422	*P*422, $P42_12$, $P4_122$, $P4_12_12$, $P4_222$, $P4_22_12$, $P4_32_12$, $P4_322$, *I*422, $I4_122$
Trigonal	3	*P*3, $P3_1$, $P3_2$, *R*3
	32	*P*312, *P*321, $P3_121$, $P3_112$, $P3_212$, $P3_221$, *R*32
Hexagonal	6	*P*6, $P6_5$, $P6_4$, $P6_3$, $P6_2$, $P6_1$
	622	*P*622, $P6_122$, $P6_222$, $P6_322$, $P6_422$, $P6_522$
Cubic	23	*P*23, *F*23, *I*23, $P2_13$, $I2_13$
	432	*P*432, $P4_132$, $P4_232$, $P4_332$, *F*432, $F4_132$, *I*432, $I4_132$

Fig. 1.9 The 65 enantiomorphic space groups.

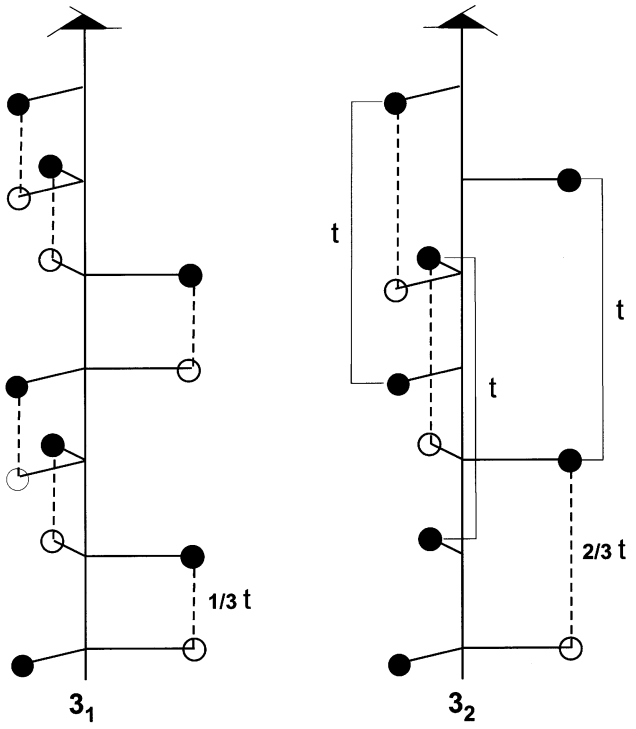


Fig. 1.10 Action of 3_1- and 3_2-screw axes.

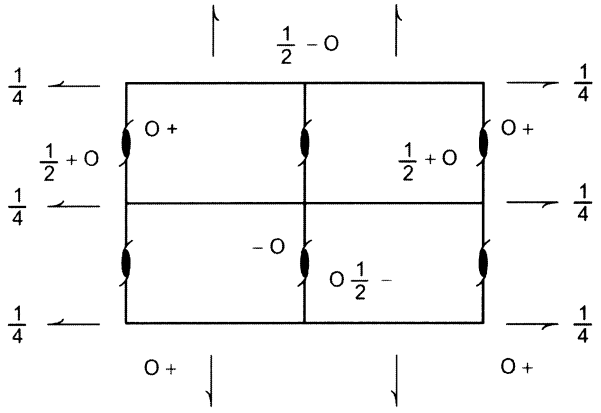


Fig. 1.11 Graphical representation of space group $P2_12_12_1$.

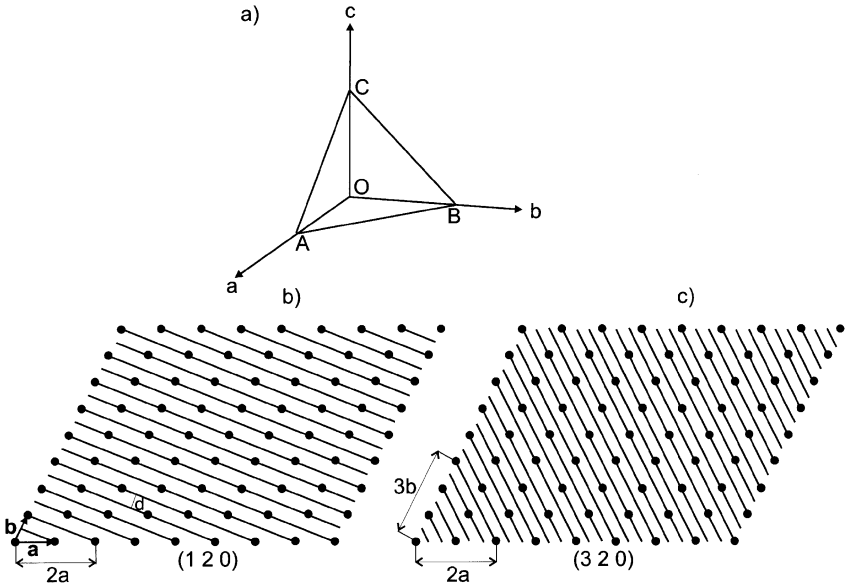


Fig. 1.12 Relationship between a crystal face and lattice planes. (a) Axes intercepts of a crystal face; (b) corresponding 2D crystal lattice with lattice plane (120); and (c) with lattice plane (320).

The N atoms contained in a crystal unit cell are at positions

$$\mathbf{r}_j = x_j\mathbf{a} + y_j\mathbf{b} + z_j\mathbf{c} \tag{1.3}$$

for the atom j with the fractional coordinates x_j, y_j, z_j, whose absolute values are between 0 and 1, and the lattice vectors **a**, **b**, and **c**. The lattice vectors follow the metric of the crystal system to which the relevant crystal belongs. As the crystallographic crystal systems are adapted to the existent crystal symmetry, the analytical form of symmetry operations adopt very concise forms. Each coordinate triplet x'_j, y'_j, z'_j is related to the symmetry operation that maps a point with coordinates x, y, z onto a point with coordinates x'_j, y'_j, z'_j. The mapping of x, y, z onto x'_j, y'_j, z'_j is given by Eq. (1.4).

$$\mathbf{x}'_j = \mathbf{W}\mathbf{x}_j + \mathbf{w} \tag{1.4}$$

with

$$\mathbf{x}_j = \begin{pmatrix} x_j \\ y_j \\ z_j \end{pmatrix}, \quad \mathbf{x}'_j = \begin{pmatrix} x'_j \\ y'_j \\ z'_j \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix}. \tag{1.5}$$

W is called the rotation part and **w** the translation part. In Figure 1.11, for example, the 2_1-axis parallel **c** going through $x = 1/4, y = 0$ maps coordinates x, y, z onto $-x + 1/2, -y, z + 1/2$ with

$$\mathbf{W} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{w} = \begin{pmatrix} 1/2 \\ 0 \\ 1/2 \end{pmatrix} \tag{1.6}$$

which can be verified by matrix multiplication.

1.2 Protein Solubility

Figure 1.13 shows a typical phase diagram illustrating the solubility properties of a macromolecule. In the labile phase crystal nucleation and growth compete, whereas in the metastable region only crystal growth appears. In the unsaturated region the crystals dissolve. The solubility of proteins is influenced by several factors, as follows.

1.2.1 Ionic Strength

A protein can be considered as a polyvalent ion, and therefore its solubility can be discussed on the basis of the Debye–Hückel theory. In aqueous solution, each ion is surrounded by an "atmosphere" of counter ions. This ionic atmosphere influences the interactions of the ion with water molecules and hence the solubility.

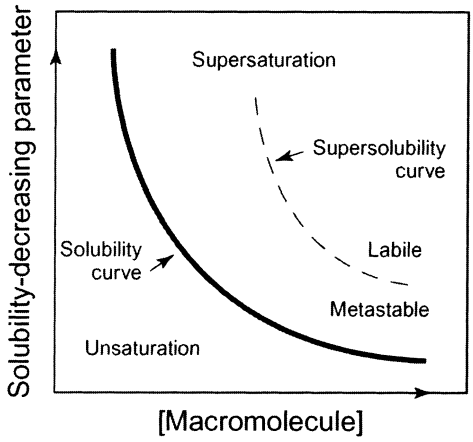


Fig. 1.13 Phase diagram illustrating the solubility properties of macromolecules. (Reproduced by permission of Academic Press, Inc., from Weber, 1997.)

1.2.1.1 **"Salting-in"**

At low ionic concentration, the "ionic atmosphere" increases the solubility as it increases the possibilities for favorable interactions with water molecules. Thus, we obtain Eqs. (1.7) and (1.8):

$$\log S - \log S_0 = \frac{AZ_+Z_-\sqrt{\mu}}{1 + aB\sqrt{\mu}} \tag{1.7}$$

$$\mu = \frac{1}{2}\sum c_j Z_j^2 \tag{1.8}$$

where μ=ionic strength, S=solubility of the salt at a given ionic strength μ, S_0=solubility of the salt in absence of the electrolyte, Z_+, Z_- the ionic charge of salt ions, A, B=constants depending on the temperature and dielectric constant, a=average diameter of ions, and c_j=concentration of the jth chemical component. Ions with higher charge are more effective for changes in solubility. Most salts and proteins are more soluble in low ionic strength than in pure water; this is termed "salting-in" (Fig. 1.14).

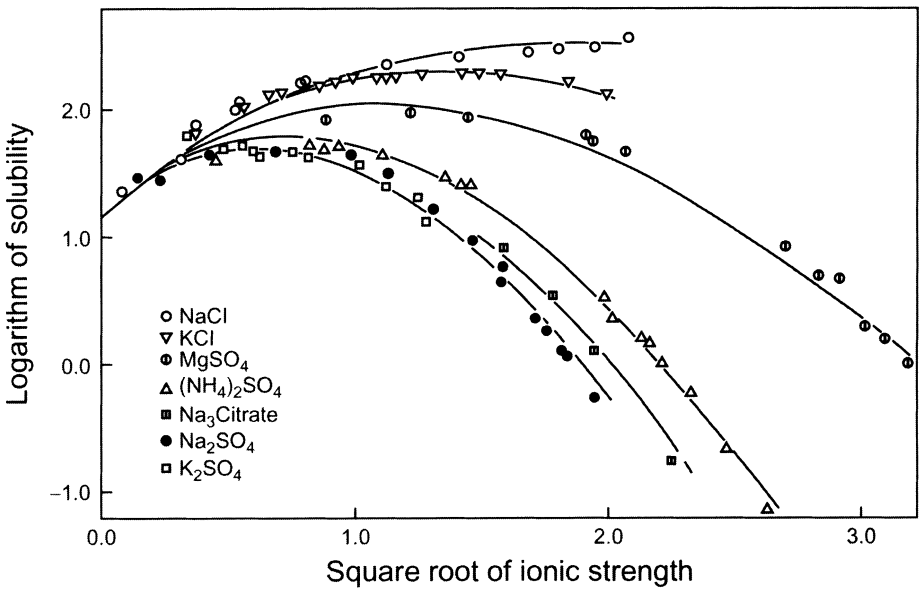


Fig. 1.14 Solubility of carboxyhemoglobin in various electrolytes at 25 °C. (Reproduced by permission of the American Society for Biochemistry and Molecular Biology, from Green, 1932.)

1.2.1.2 "Salting-out"

At higher ionic strength the ions compete for the surrounding water, and consequently the water molecules are taken away from the dissolved agent and the solubility decreases according to Eq. (1.9):

$$\log S - \log S_0 = \frac{AZ_+Z_-\sqrt{\mu}}{1 + aB\sqrt{\mu}} - K_s\mu \tag{1.9}$$

The term $K_s\mu$ predominates at high ionic strengths, which means that "salting-out" is then proportional to the ionic strength (Fig. 1.14). In a medium with low ionic strength, the solubility of a protein can be decreased by increasing or decreasing the salt concentration. Salts with small, highly charged ions are more effective than those with large, lowly charged ions. Ammonium sulfate is often used because of its high solubility.

1.2.2 pH and Counterions

The more soluble a protein, the larger is its net charge, with the minimum solubility being found at the isoelectric point. The net charge is zero, and hence the packing in the solid state (in the crystal) is possible owing to electrostatic interactions without the accumulation of a net charge of high energy. All "salting-out" curves are parallel, K_s remains constant, and S_0 varies with pH (Fig. 1.15a and b). In some cases the isoelectric point is different at low and high ionic strengths, owing to interactions of the protein with counterions which can cause a net charge at the pH of the isoelectric point.

1.2.3 Temperature

Many factors governing protein solubility are temperature-dependent. The dielectric constant decreases with increasing temperature. In the solution energy, $\Delta G = \Delta H - T\Delta S$, the entropy term has an increasing influence with increasing temperature. The temperature coefficient of the solubility depends on other conditions (ionic strength, presence of organic solvents, etc.). At high ionic strength most proteins are less soluble at 25 °C than at 4 °C – that is, the temperature coefficient is negative. The opposite is valid for low ionic strength.

1.2.4 Organic Solvents

The presence of organic solvents leads to a decrease in the dielectric constant. This causes an augmentation of the electric attraction between opposite charges on the surface of the protein molecule, and hence to a reduction in solubility. In general, the solubility of a protein is reduced in the presence of an organic

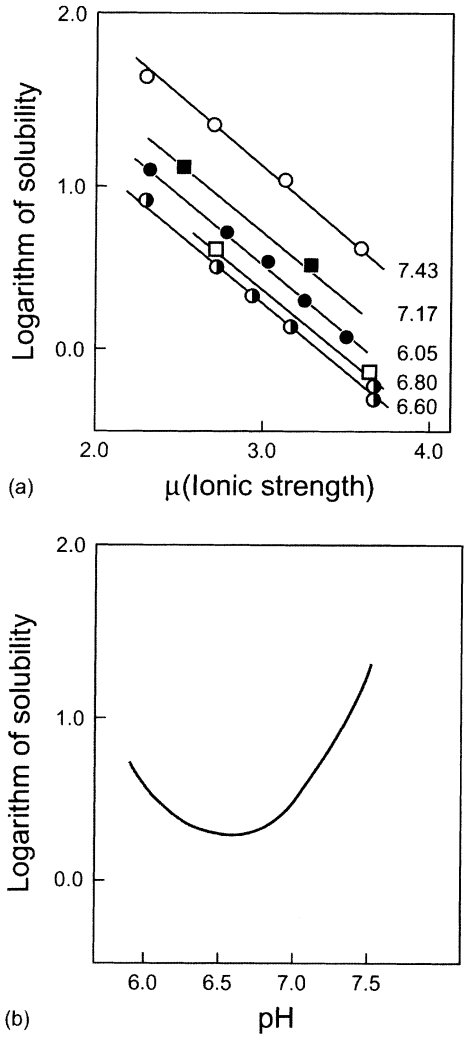


Fig. 1.15 (a) Solubility of hemoglobin at different pH values in concentrated phosphate buffers; (b) extracted from (a). (Reproduced by permission of the American Society for Biochemistry and Molecular Biology, from Green, 1931.)

solvent if the temperature decreases. Often, organic solvents denature proteins, and consequently one should work at low temperatures.

1.3 Experimental Techniques

The whole field of macromolecular crystallography has been excellently reviewed in Volumes 114 and 115 and Volumes 276 and 277 of *Methods in Enzymology*. A collection of review articles concerning the theory and practice of crys-

tallization of biomacromolecules is provided in Part A of Carter and Sweet (1997).

A protein preparation to be used in crystallization studies should be "pure" or "homogeneous" at a level that established chromatographic methods are providing (protein content ≥95%). Furthermore, it should meet the requirements of "structural homogeneity". These requirements can be enumerated as follows. It is first necessary to prepare the protein in an isotypically pure state free from other cellular proteins. It may then be necessary to maintain the homogeneity of the protein preparation against covalent modification during crystallization by adding inhibitors of sulfhydryl group oxidation, proteolysis and the action of reactive metals. It may be necessary to suppress the slow denaturation/aggregation of the protein and to restrict its conformational flexibility to reduce the entropic barrier to crystallization presented by extensive conformational flexibility. For the crystallization of biomacromolecules, a broad spectrum of crystallization techniques exists, the most common of which are described here.

1.3.1 Batch Crystallization

This is the oldest and simplest method (see Fig. 1.16 a). In batch experiments, vials containing supersaturated protein solutions are sealed and left undisturbed. In microbatch methods, a small (2–10 μL) droplet containing both protein and precipitant is immersed in an inert oil which prevents droplet evaporation. In the case that ideal conditions for nucleation and growth are different, it is useful to undertake the separate optimization of these processes. This can be done by seeding – a technique where crystals are transferred from nucleation conditions to those that will support only growth (Fig. 1.16 b). For macroseeding, a single crystal is transferred to an etching solution, then to a solution of optimal growth. In microseeding experiments, a solution containing many small seed crystals, occasionally obtained by grinding a larger crystal, is transferred to a crystal growth solution.

1.3.2 Vapor Diffusion

Crystallization by vapor diffusion is depicted in Figure 1.17 a. Here, unsaturated precipitant-containing protein solutions are suspended over a reservoir. Vapor equilibration of the droplet and reservoir causes the protein solution to reach a supersaturation level where nucleation and initial crystal growth occur. Changes in soluble protein concentration in the droplet are likely to decrease supersaturation over the time course of the experiment. The vapor diffusion technique can be carried out as either a hanging drop or sitting drop method.

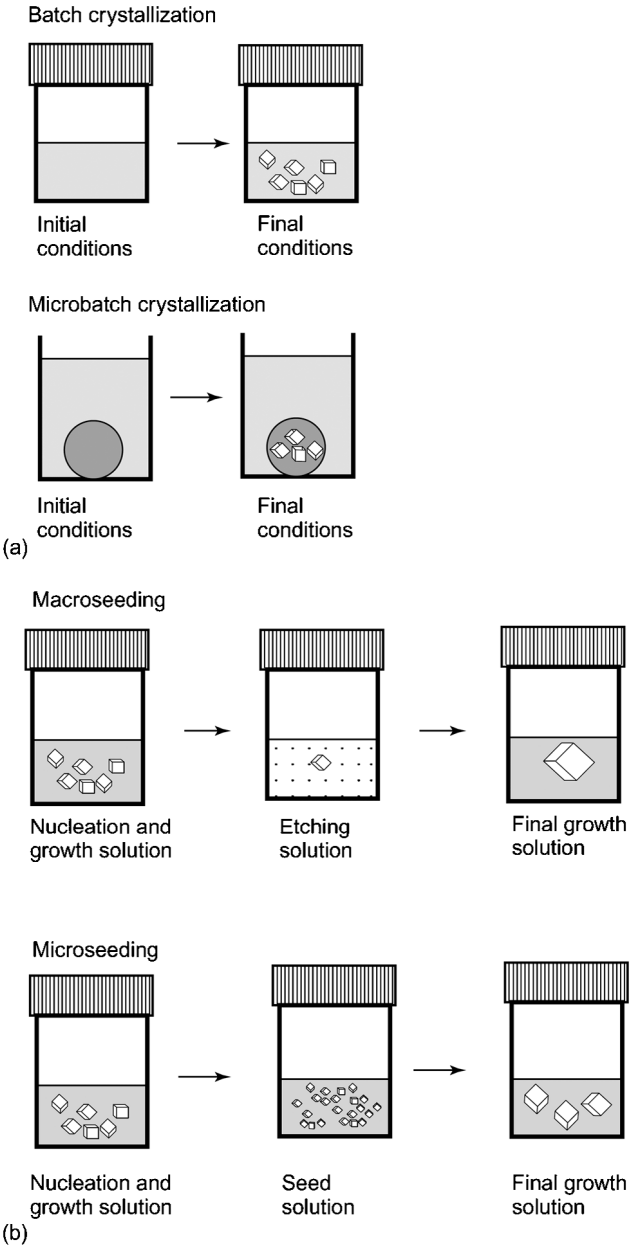


Fig. 1.16 Schematic presentation of (a) batch crystallization and (b) seeding techniques. (Reproduced by permission of Academic Press, Inc., from Weber, 1997.)

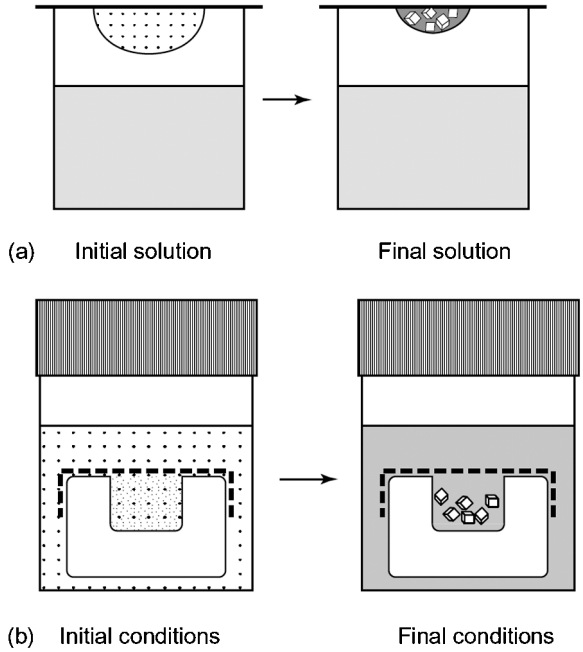


Fig. 1.17 Schematic representation of (a) vapor diffusion and (b) dialysis. (Reproduced by permission of Academic Press, Inc., from Weber, 1997.)

1.3.3
Crystallization by Dialysis

In crystallization by dialysis, the macromolecular concentration remains constant, as in batch methods (Fig. 1.17b) because the molecules are forced to stay in a fixed volume. The solution composition is changed by diffusion of low-molecular-weight components through a semipermeable membrane. The advantage of dialysis is that the precipitating solution can be easily changed. Dialysis is also uniquely suited to crystallizations at low ionic strength and in the presence of volatile reagents such as alcohols.

1.4
Crystallization Screenings

Screening schemes have been developed which change the most common parameters of this multiparameter problem, such as protein concentration, the nature and concentration of the precipitant, pH, and temperature. Each screening can be extended by adding specific additives in low concentrations that affect the crystallization. Sparse matrix crystallization screens are widely applied. The

sparse matrix formulation allows the efficient screening of a broad range of the most popular and effective salts (e.g., ammonium sulfate, sodium and potassium phosphate, sodium citrate, sodium acetate, lithium sulfate), polymers [e.g., poly(ethyleneglycol) (PEG) of different molecular masses (from 400 to 8000)], and organic solvents [e.g., 2,4-methylpentanediol (MPD), 2-propanol, ethanol) versus a wide range of pH. Another approach is the systematic screening of the statistically most successful precipitants. A single precipitant is screened at four unique concentrations versus seven precise levels of pH between 4 and 10. Such grid screens can be carried out with ammonium sulfate, PEG 600, MPD, and PEG 6000 in the presence of 1.0 M lithium chloride or sodium chloride. For the crystallization of membrane proteins (see Michel, 1991) for each detergent which is necessary to solubilize the membrane protein, a whole grid screen or sparse matrix screen must be constructed. In principle, all three techniques can be applied for the different screening schemes, but in the most part the vapor diffusion technique is applied because it is easy to use and the protein consumption is low. For a typical broad screening, about 2 mg of protein is sufficient. Chryschem plates (sitting drop) or Linbro plates (hanging drops) may be used for the vapor diffusion crystallization screening experiments. Once crystals have been obtained, their size and quality can be optimized by additional fine screens around the observed crystallization conditions. There are no general rules to indicate which method should be used to crystallize which type of protein; however, suggestions for crystallization conditions to be tested can be obtained from the Biological Macromolecule Crystallization Database (Gilliland et al., 1994; http://xpdb.nist.gov:8080/bmcd/bmcd.html).

1.5 High-Throughput Crystallization, Imaging, and Analysis

During recent years, the sequencing of whole genomes from bacteria to higher organisms, including man, has opened up the systematic determination of their gene products. Today, this new field is known as “structural genomics” or “structural proteomics”. Structural genomics represents not only the structure determination of gene products, by using the old approach of structural biology, one target, one researcher, but also comprises the creation and application of high-throughput techniques. Unfortunately, these major efforts can be managed only by larger consortia, and several such set-ups have been established in the USA, Japan, and Europe. A complete list can be found on the Internet under http://sg.pdb.org/target_centers.html. The automation includes the whole workflow in protein structure determination from cloning, expression, purification, quality assessment, crystallization, imaging, X-ray data collection, and structure analysis.

The focus of the following section is on high-throughput crystallization, crystal imaging, and image analysis. Today, crystallization robots have been developed that not only automate the crystallization set-ups but also reduce the vol-

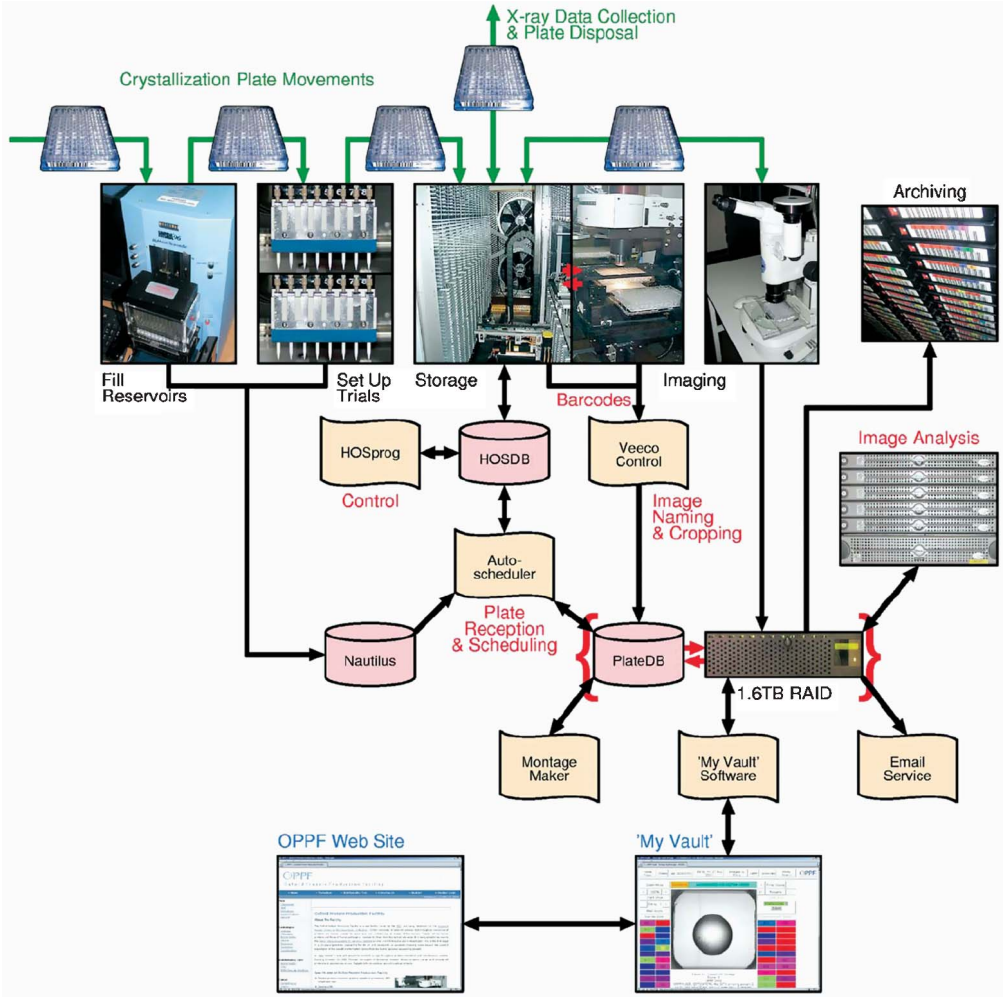


Fig. 1.18 Scheme of the components and workflow of the OPPF high-throughput crystallization facility. Green arrows show the transfer of 96-well crystallization plates between robots. The flow of images and control data is shown by black arrows. Databases are indicated by pale red "disk cylinders" and specific sections of the control software are represented by orange "paper". The method of interaction with the web interface is indicated by sample web pages. (Reproduced by permission of Elsevier Ltd., from Mayo et al., 2005.)

ume of the dispensed protein drops from µL quantities to 50 nL. This dramatically increases the number of screening conditions with the same amount of available protein. Several facilities have been set up, which have completely automated the liquid and protein dispensing, the plate storage, imaging, and image analysis. A number of these systems are now also available commercially.

In principle, all systems contain the same components, and in the following section a typical large-scale facility installed at the Oxford Protein Production Facility (OPPF) is described in more detail (Fig. 1.18). The initial crystallization screening uses a panel of 480 conditions selected from standard (commercially available) crystallization kits. The kits are reformatted into 96-deep well "Master blocks" by a Qiagen Biorobot 8000. Pre-barcoded 96-well crystallization plates (Greiner Bio-One Ltd, UK) are used for the trials, the precipitant being transferred from the master blocks to the reservoirs using a Hydra-96 microdispenser (Matrix Technologies Ltd, UK). The barcode of the plate is then read and transferred to the LIMS (Laboratory Information Management System). The plate is then placed on a Cartesian Technologies Microsys MIC400 (Genomic Solutions Ltd, UK) where a 100-nL drop of protein solution is placed on the central position of each crystallization shelf and mixed with 100 nL of the corresponding reservoir. The pipetted plates are sealed and stored in an automated storage vault (The Automated Partnership Ltd, UK). Imaging is performed using an Oasis 1700 automatic imaging system (Veeco, UK), which is housed in an annex to the storage vault. Plates can be picked by a robot arm in the storage vault and transferred to the imaging system controlled by the LIMS. In this way, a 96-well plate can be imaged in 40 s. The digitized images are transferred to a RAID storage system, and each well image is classified using the York crystal image analysis software (Wilson, 2004). The program assigns different scores to the images, ranging from 0 for insignificant objects, such as those due to shadows at the edge of the drop, to 6 for good single crystals. Figure 1.18 also shows the components and arrangement of the computer hardware and LIMS.

References

Carter, C.W. Jr, Sweet, R.M. (Eds.), *Macromolecular Crystallography*, Part A, *Methods Enzymol*. **1997**, *276*, 1–700.

Carter, C.W. Jr, Sweet, R.M. (Eds.), *Macromolecular Crystallography*, Part B, *Methods Enzymol*. **1997**, *277*, 1–664.

Gilliland, G.L., Tung, M., Balkeslee, D.M., Ladner, J.E. *Acta Crystallogr.* **1994**, *D50*, 408–413.

Green, A.A. *J. Biol. Chem*. **1931**, *93*, 495–516.

Green, A.A. *J. Biol. Chem*. **1932**, *95*, 47–66.

Hahn, T. (Ed.), *International Tables for Crystallography, Volume A*, Springer, Dordrecht, **2005**.

Mayo, C.J., Diprose, J.M., Walter, T.S., Bery, I.M., Wilson, J., Owens, R.J., Jones, E.Y., Harlos, K., Stuart, D.I., Esnouf, R.M. *Structure* **2005**, *13*, 175–182.

Michel, H. (Ed.), *Crystallization of Membrane Proteins*, CRC Press, Boca Raton, **1991**.

Weber, P.C. *Methods Enzymol*. **1997**, *276*, 13–23.

Wilson, J. *Cryst. Rev.* **2004**, *10*, 73–84.

2
Experimental Techniques

2.1
X-Ray Sources

2.1.1
Conventional X-Ray Generators

X-rays are produced when a beam of high-energy electrons, which have been accelerated through a voltage V in a vacuum, hit a target. An X-ray tube run at voltage V will emit a continuous X-ray spectrum with a minimum wavelength given by Eq. (2.1):

$$\lambda_{\min} = \frac{hc}{ev} = \frac{12398}{V} \tag{2.1}$$

with λ in Angströms ($1\ \text{Å} = 10^{-10}$ m) and V in volts. The critical voltage, V_0, which is required to excite the characteristic line of a particular element, can be calculated from the corresponding wavelength for the appropriate absorption edge. For the copper absorption edge $\lambda_{ae} = 1.380$ Å. Hence:

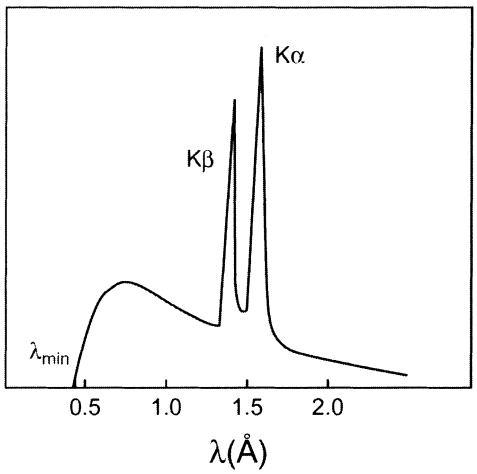


Fig. 2.1 X-ray spectrum emitted from a copper anode. This shows the continuous "Bremsspektrum" starting at λ_{min} and the two characteristic copper lines λ Kα = 1.5418 Å (superposition of λ Kα_1 = 1.5405 Å and λ Kα_2 = 1.5443 Å) and λ Kβ = 1.3922 Å.

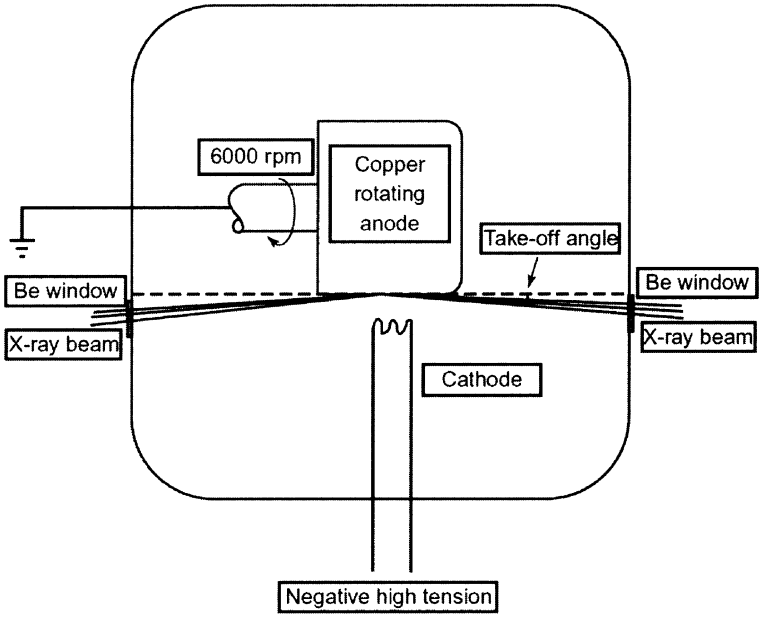


Fig. 2.2 Schematic drawing of a rotating anode tube. The take-off angle is close to 4°. For copper, the tube is normally operated at 50 kV high tension and 100 mA cathode current.

$$V_0 = \frac{12398}{\lambda_{ae}} = 8.98 \text{ kV} \tag{2.2}$$

Provided that $V > V_0$, the characteristic line spectra will be produced (Fig. 2.1). The oldest and cheapest X-ray sources are sealed X-ray tubes, where the cathode and anode are situated under vacuum in a sealed glass tube and the heat generated at the anode is removed by a water-cooling system. For the generation of higher intensities, as needed in protein crystallography, it is necessary to use a rotating anode (Fig. 2.2). Here, the anode is rotated, which allows a higher power loading at the focal spot. In protein crystallography copper targets are usually taken. The used take-off angle is close to 4°, which results in apparent focal spot sizes of about 0.3×0.3 mm.

2.1.2 Synchrotron Radiation

As electrically charged particles such as electrons or positrons of high energy are kept under the influence of magnetic fields and travel in a pseudocircular trajectory, synchrotron radiation is emitted. This can be used in many different types of experiments (for a comprehensive discussion of synchrotron radiation in macromolecular crystallography, see Helliwell, 1992). For relativistic electrons

with energy E, the electromagnetic radiation is compressed into a fan-shaped beam tangential to the orbit with a vertical opening angle $\Psi \cong mc^2/E$, i.e., 0.1 mrad for $E=5$ GeV (Fig. 2.3). As this fan rotates with the circulating electrons or positrons, a stationary observer will see n flashes of radiation every $2\pi R/c$ s, the duration of each flash being less than 1 ns. The spectral distribution of synchrotron radiation extends from the infrared to the X-ray region (Fig. 2.4). An important parameter is the median of the distribution of power over the spectral region, called the "critical photon energy" E_c, which divides the power spectrum into two equal parts. Taking the wavelength λ instead of the photon energy E, we obtain for the critical wavelength

$$\lambda_c = 18.64/(BE^2)\,, \tag{2.3}$$

where B (=3.34 E/R) is the magnetic bending fielding T, E is in GeV, and R in meters.

The particles are injected into the storage ring directly from a linear accelerator or through a booster ring (Fig. 2.5), and circulate in a high vacuum for several hours at a relative constant energy. In order to keep the bunched particles traveling in a near-circular path, a lattice of bending magnets is set up around the storage ring. As the particle beam traverses each magnet, the path of the beam is altered, and synchrotron radiation is emitted. The loss of energy of the particle beam is compensated by an oscillating radiofrequency (RF) electric field at each cycle. Synchrotron radiation is highly polarized. In an ideal ring, where

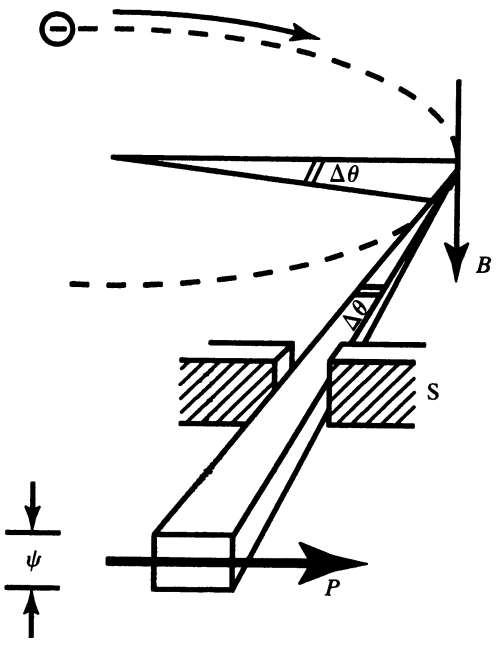


Fig. 2.3 Synchrotron radiation emitted by a relativistic electron traveling in a curved trajectory. B is the magnetic field perpendicular to the plane of the electron orbit; ψ is the natural opening angle in the vertical plane; P is the direction of polarization. The slit, S, defines the length of the arc of angle, $\Delta\theta$, from which the radiation is taken. (From Buras and Tazzari, 1984.)

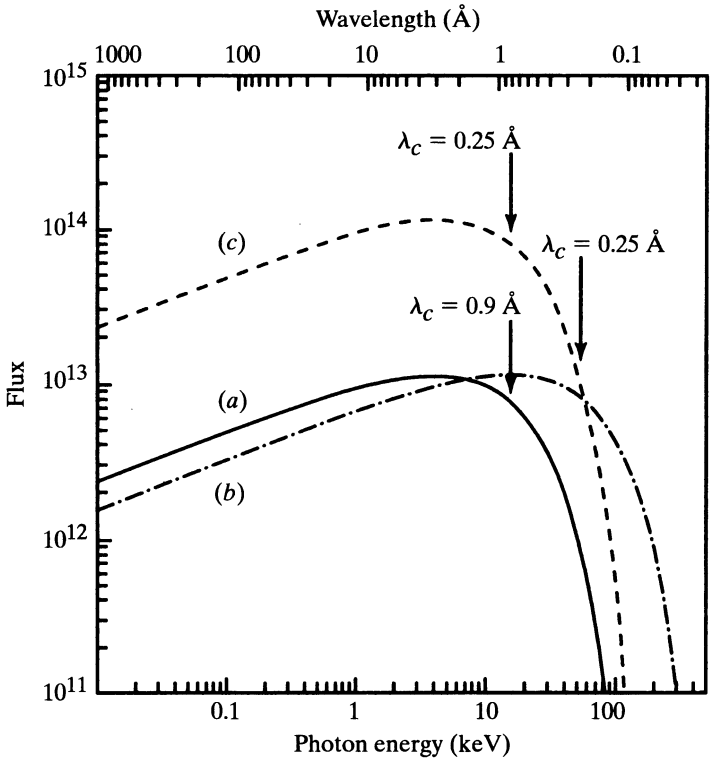


Fig. 2.4 Spectral distribution and critical wavelengths for: (a) a dipole magnet; (b) a wavelength shifter; and (c) a multipole wiggler at the ESRF. (From Buras and Tazzari, 1984.)

all electrons are parallel to one another in a central orbit, the radiation in the orbital plane is linearly polarized, with the electric vector lying in this plane (Fig. 2.3). Outside this plane, the radiation is elliptically polarized.

The synchrotron radiation can be channeled through different beamlines for use in research. Other types of magnets – insertion devices called "wigglers" and "undulators" – can be assembled in the storage ring, which is in practice not a circle. These have a zero magnetic field integral and may be inserted into the straight sections (see Fig. 2.5). Unlike the bending magnets – the primary purpose of which is to maintain the circular trajectory – wigglers and undulators are used to increase the intensity of the emitted radiation. Bending magnets and wigglers cause a continuous spectrum of radiation.

A *wiggler* consists of one or more dipole magnets with alternating magnetic field directions aligned transverse to the orbit. The critical wavelength can thus be shifted towards shorter values because the bending radius can be decreased over a short section. Such a device is called a "wavelength shifter". A series of N dipole magnets constitutes a multipole wiggler. The electron trajectory in such a

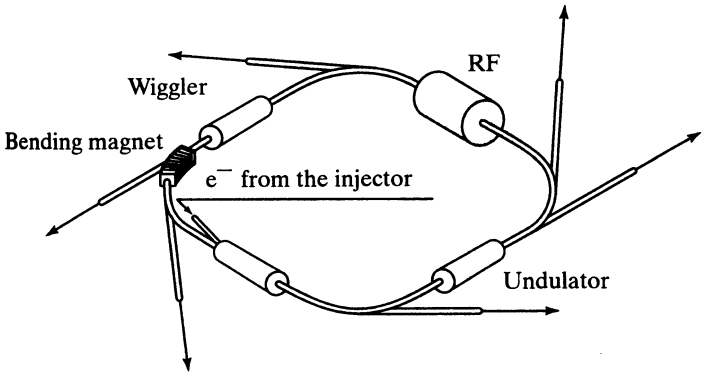


Fig. 2.5 The main components of a dedicated electron storage-ring synchrotron-radiation source. For clarity, only one bending magnet is shown. (From Buras and Tazzari, 1984.)

device is shown in Figure 2.6. The flux of N dipoles adds up to N times the flux of a single dipole.

A multipole wiggler becomes an *undulator* if the magnet poles have a short period and $2a << \psi$. Interference takes place between radiation of wavelength λ_0 emitted at two points λ_0 apart on the electron trajectory (Fig. 2.6). The spectrum at an angle θ to the axis observed through a pinhole has a peak at a specific wavelength and a few harmonics.

The importance of synchrotron radiation for macromolecular crystallography lies in the high brilliance (photons s^{-1} $mrad^{-2}$ mm^{-2} per $\Delta\lambda/\lambda$; that is, how small is the source and how well collimated are the X-rays?) of the beam, the high intensity, and the tunability of the wavelength in the relevant range from 0.5 to 3.0 Å. The time structure of the beam is of interest for time-resolved crystallography (Moffat, 1998). The particles circulate in bunches with widths of 50 to 150 ps, and repeat every few microseconds.

About 15 synchrotron radiation facilities equipped with beamlines for macromolecular crystallography are available worldwide, and are operated at energies from about 1.5 to 6–8 GeV for third-generation machines. An aerial view of the European Synchrotron Radiation Facility (ESRF) in Grenoble, a third-generation

Fig. 2.6 Electron trajectory within a multipole wiggler or undulator. λ_0 is the spatial period, a is the maximum deflection angle, and θ is the observation angle. (From Buras and Tazzari, 1984.)

Fig. 2.7 An aerial view of the European Synchrotron Radiation Facility (ESRF) in Grenoble. (Illustration courtesy of ESRF.)

machine, is shown in Figure 2.7. The ESRF storage ring is operated at 6 GeV and has a circumference of 844.39 m. Its critical wavelength, λ_c, is 0.6 Å.

2.1.3 Monochromators

In the majority of applied diffraction techniques, monochromatic X-rays are used. Therefore, the emitted white radiation of X-rays must be further monochromatized. With copper Kα radiation generated by a sealed or rotating anode tube, the Kβ radiation can be removed with a nickel filter. However, much better results can be achieved with a monochromator. The simplest monochromator is a piece of a graphite crystal which reflects the copper Kα radiation at a Bragg angle of 13.1° and a glancing angle of 26.2°. Improved beam focusing is obtained by a double mirror system, where total reflection at grazing angles is used. This technique was first introduced by Kirkpatrick and Baez (1948), and is shown schematically in Figure 2.8. In this geometry, one reflector focuses in one dimension. Focusing in the second dimension is obtained by a second reflector downstream perpendicular to the first one. In the commercially available version (Molecular Structure Corporation, The Woodlands, TX, USA), the mirror assembly is composed of two bent nickel-coated glass optical flats, each with translation, rotation and slit components housed in a helium gas flashed chamber. The prototype and basic theory in the use of this system were discussed in detail by Phillips and Rayment (1985).

Recently, total reflection coatings have been replaced by appropriate laterally graded multilayers. X-rays from a divergent source will strike an aspherically curved surface at different angles of incidence. The graded d-spacing allows

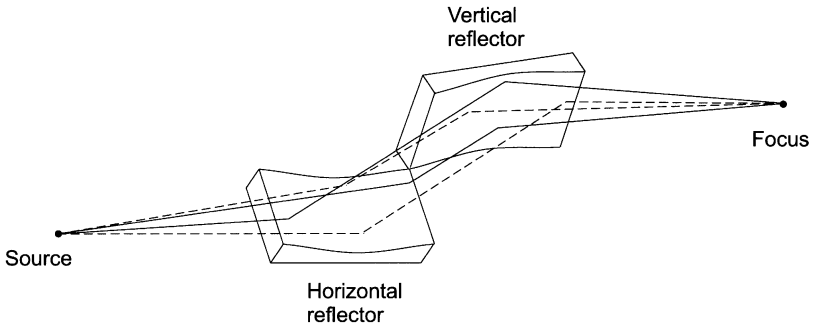


Fig. 2.8 Double mirror arrangement according to Kirkpatrick and Baez (1948).

every point along the optic to satisfy the Bragg reflection condition. Thus, this optic is capable of transforming a divergent beam into either a parallel or focused beam. Figure 2.9 explains the mode of action of a focusing multilayer optic. The smaller layer distances at the entrance side of the multilayer block reflect at larger Bragg angles than the larger ones.

It is an interesting point that the angles are a factor of 3 to 4 larger in Bragg diffraction compared to total reflection, which makes it feasible to position both reflectors side-by-side. This has been realized in the Confocal Max-Flux^{TM} Optic (Osmic Inc., The Woodlands, TX, USA). Both reflectors are part of the same block but are positioned at right-angles to each other (Fig. 2.10). The operation is similar to that of the Kirkpatrick–Baez schemes, in that each reflector focuses the beam in one direction only, but the situation is slightly more complicated. The X-rays can pass through the optic in two different ways: either by first reflecting from reflector 1 and afterwards from reflector 2 (full line in Fig. 2.10), or from reflector 2 to reflector 1 (dashed line). In general, the width of the reflectors will be larger than required, and consequently reflection will take place

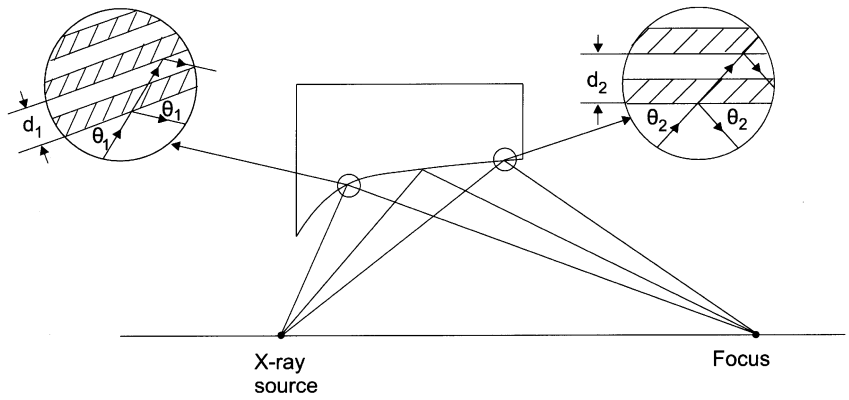


Fig. 2.9 Mode of action of a focusing graded multilayer reflector for X-rays.

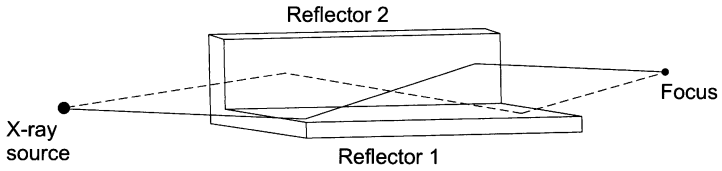


Fig. 2.10 Beam path in the Confocal Max-Flux™ Optic (Osmic Inc., The Woodlands, TX, USA).

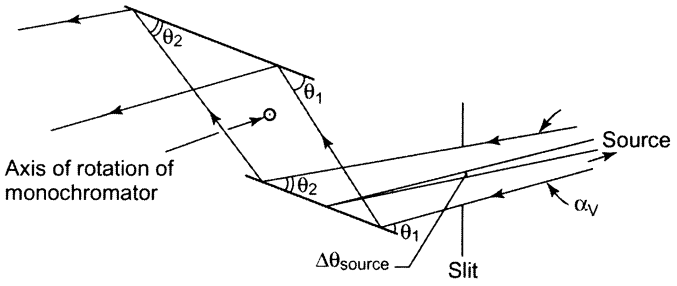


Fig. 2.11 Schematic drawing of a double monochromator system. (Reproduced by permission of Cambridge University Press, from Helliwell, 1992.)

only near the intersection of the two reflectors. Such monochromators are now used in many in-house protein crystallography X-ray diffraction machines.

For synchrotron radiation, with its much higher intensity, germanium or silicon single crystals can be applied as monochromators which filter out a bandwidth of $\delta\lambda/\lambda$ from 10^{-4} to 10^{-5}, two orders of magnitude smaller than with graphite. Single or double monochromators can be used which are either flat or bent. The bent monochromators have the advantage that they simultaneously focus the beam. The double monochromator (Fig. 2.11) has the advantage that the emergent monochromatic beam is parallel to, and only slightly displaced from, the incident synchrotron radiation beam. This makes necessary only small adjustments of the X-ray optics and detector arrangement when it is tuned to another wavelength compared with a single monochromator, where the whole X-ray diffraction assembly must be moved. Common beamline optic modes are shown in Figure 2.12.

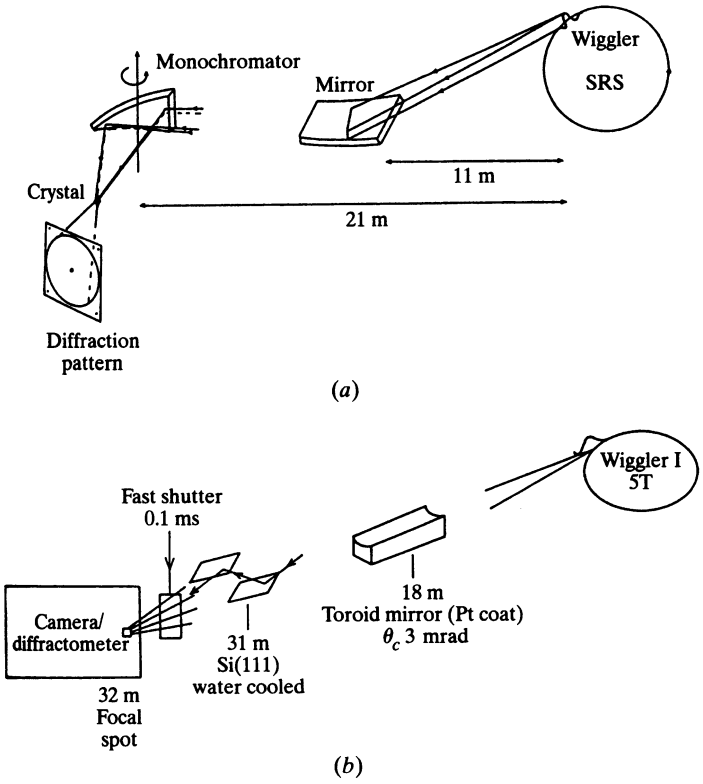


Fig. 2.12 Common beamline optic modes. (a) Horizontally focusing cylindrical monochromator and vertical focusing mirror (shown here for station 9.6 at the Daresbury SRS facility). (b) Rapidly tuning double-crystal monochromator and point-focusing toroid mirror (shown here for station 9.5 at the Daresbury SRS facility). (Reproduced by permission of Kluwer Academic Publishers, from Helliwell, 2001.)

2.2
Detectors

2.2.1
General Components of an X-Ray Diffraction Experiment

A principal arrangement for a macromolecular X-ray diffraction experiment is depicted in Figure 2.13. The primary beam leaves the X-ray source and passes the X-ray optics, which may be a simple collimator or the various types of monochromators or mirror systems described above and terminated with a collimator. The crystal is mounted on a goniometer head, either in a quartz capillary or in a cryo-loop shock-frozen at low temperature. The goniometer head is attached to a device which can perform spatial movements of the crystal around the center of the crystal. The simplest kind of such a movement is rotation of

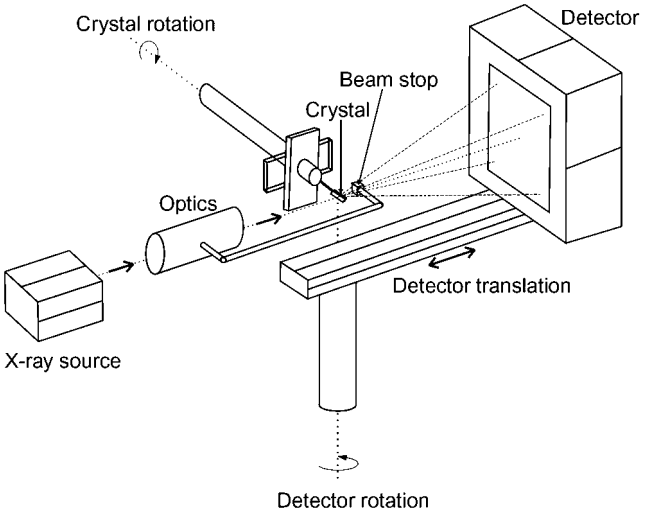


Fig. 2.13 The principal arrangement for a macromolecular X-ray diffraction experiment.

the crystal about a spindle axis, as indicated in Figure 2.13. This device can be a multiple axis goniostat (2–4 axes) which allows the crystal to be brought into any spatial orientation around its center. The X-ray detector which registers the diffracted intensities is mounted on a device which permits the translation and rotation of the detector. If the active area of the detector is large enough to collect all generated diffracted beams at a given wavelength, then detector rotation is not necessary and the detector is arranged normal to the primary beam. A small piece of lead is placed in the path of the primary beam just behind the crystal to prevent damage to the detector and superfluous gas scattering.

In the past, the classical detectors in macromolecular crystallography have been photographic films and single-photon counters. The photographic films were used on specially designed X-ray cameras, and the single-photon counters on four-circle diffractometers. The main disadvantage of these detectors was their low sensitivity, and with films it was the limited dynamic range (1 : 200). Over the past 15 years, however, powerful detectors have been developed which will be discussed briefly here. These new detectors have almost completely replaced photographic films and single-photon counters.

2.2.2 Image Plates

An image plate (IP) consists of a support (either a flexible plastic plate or a metal base) coated with a photostimulable phosphor (150 μm thickness) and a protective layer (10 μm). The photostimulable phosphor is a mixture of very thin crystals of, for example, $BaF(Br,I):Eu^{2+}$ and an organic binder. This phosphor can store a fraction of the absorbed X-ray energy by electrons trapped in color

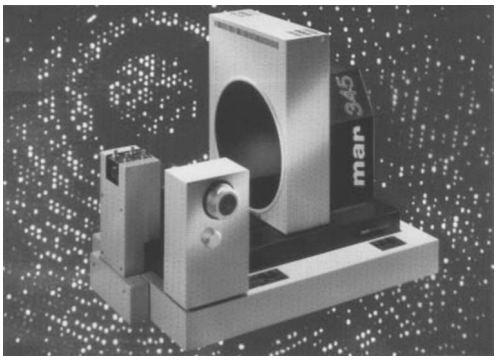


Fig. 2.14 The Mar 345 IP diffractometer system from Mar Research, Norderstedt, Germany. The circular plate rotates for scanning, and the laser is moved along a radial line. (Illustration courtesy of Mar Research.)

centers. It emits photostimulated luminescence, the intensity of which is proportional to the absorbed X-ray intensity, when later stimulated by visible light. The wavelength of the photostimulated luminescence ($\lambda \approx 390$ nm) is reasonably separated from that of the stimulating light ($\lambda \approx 633$ nm, in practice a red laser), allowing it to be collected by a conventional high quantum efficiency photomultiplier tube. The output of the photomultiplier is amplified and converted to a digital image, which can be processed by a computer. The residual image on the IP can be erased completely by irradiation with visible light, to allow repeated use.

IPs have several excellent performance characteristics as integrating X-ray area detectors that make them well suited for X-ray diffraction. The sensitivity is at least 10-fold higher than for X-ray films, and the dynamic range is much broader ($1:10^4$–10^5). One important point for synchrotron radiation is their high sensitivity at shorter wavelengths (e.g., 0.65 Å). However, a disadvantage is the relatively long readout times for each exposure (from 45 s to several minutes). IP diffractomer systems are available commercially from several companies, and all systems operate reliably and deliver good-quality data. A photograph of the newest IP system produced by Mar Research (Norderstedt, Germany) is shown in Figure 2.14.

2.2.3
Gas Proportional Detectors

As X-ray counters, gas proportional detectors provide unrivaled dynamic range and sensitivity for photons in the range which is important for macromolecular crystallography (for a review, see Kahn and Fourme, 1997). The classical gas proportional detector is a multiwire proportional chamber (MWPC), widely used as an in-house detector with conventional X-ray sources. Two MWPC diffractometer systems are commercially available. Gas proportional detectors use as a first step the absorption of an X-ray photon in a gas mixture high in xenon or argon. This photoabsorption produces one electron–ion pair, the total energy of which is simply the energy of the initial X-ray photon. The ion returns to its

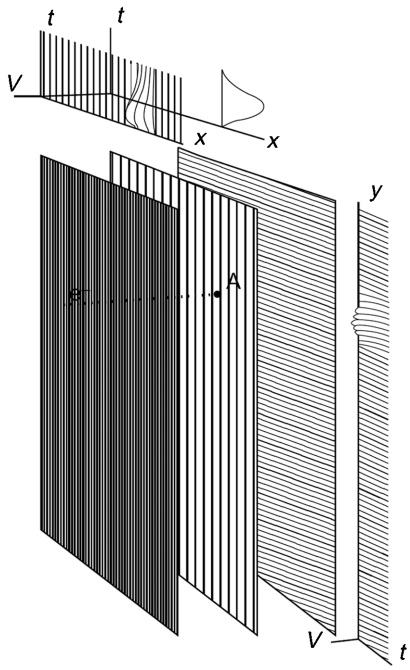


Fig. 2.15 Expanded view of a multiwire proportional chamber (MWPC) showing the anode plane sandwiched between the two cathode planes. A is the position of the avalanche. The centers of the induced charge distributions are used to determine the coordinates, *x* and *y*, of the avalanche. (Reproduced by permission of Academic Press, Inc., from Kahn and Fourme, 1997.)

neutral state either by the emission of Auger electrons, or by fluorescence. Since the kinetic energy of these first electrons is far greater than the energy of the first ionization level of the xenon or argon atoms, fast collisions with atoms (or molecules) in the gas very quickly produce a cascade of new electron–ion pairs in a small region extending over a few hundred micrometers around the conversion point. The total number of primary electrons produced during this process is proportional to the energy of the absorbed X-ray photon, and is thus a few hundred for $\sim$10 keV photons. These primary electrons then drift to the nearest anode wire where an ionization avalanche of 10 000–1 000 000 as many ion pairs results. The motion of the charged particles in this avalanche (chiefly the motion of the heavy positive ions away from the anode wire) causes a negative-going pulse on the anode wire and positive-going pulses on a few of the nearest wires in the back (cathode) wire plane (see Fig. 2.15).

The disadvantages of the MWPC detector are the limited counting rate due to the build-up of charges in the chamber, together with limitations in the readout electronics and the lower sensitivity at shorter wavelengths. This makes the application of MWPCs with synchrotron radiation poorly effective.

2.2.4
Charge-Coupled Device-Based Detectors

A remarkable development for the use with synchrotron radiation is the design and construction of charge-coupled device (CCD) detectors (for a review, see Westbrook and Naday, 1997). CCDs were developed originally as memory devices, but the observation of localized light-induced charge accumulation in CCDs quickly led to their development as imaging sensors. These CCD detectors are integrating detectors like the conventional X-ray-sensitive film, IPs and analog electronic detectors using either silicon-intensified target (SIT) or CCD sensors. Integrating detectors have virtually no upper rate limits because they measure the total energy deposited during the integration period (although individual pixels may become saturated if the signal exceeds its storage capacity).

The first commercially available analog electronic detector was the fast area television (FAST) detector produced by Enraf-Nonius (Delft, The Netherlands). This detector contained a SIT vidicon camera as an electronically readable sensor. The SIT vidicon exhibits higher noise than CCDs, which have therefore replaced SIT sensors during the past few years. Because of their high intrinsic noise, detectors with SIT vidicon sensors need an analog image-amplification stage, and this limits the overall performance of such detectors. Several CCD detector systems have also been developed that incorporate image intensification. The most important development in detector design for macromolecular crystallography has been the incorporation of scientific-grade CCD sensors into instruments with no image intensifier. These detector designs are based on direct contact between the CCD and a fiber-optic taper. There are several commercial systems available based on this construction (Hamlin Detector; Mar Research, Norderstedt, Germany).

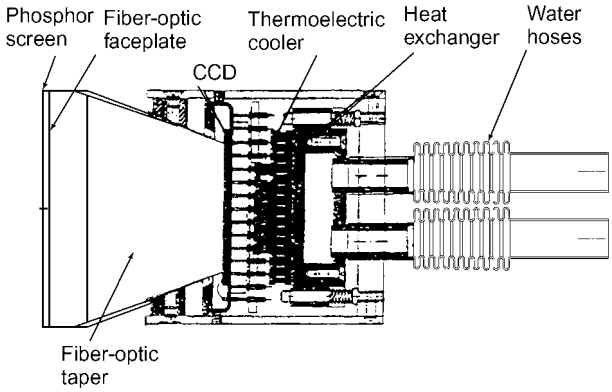


Fig. 2.16 Schematic representation of a CCD/taper detector. (Reproduced by permission of Academic Press, Inc., from Westbrook and Naday, 1997.)

A schematic representation of such a detector is shown in Figure 2.16. An X-ray phosphor (commonly $Gd_2O_2S:Tb$) is attached to a fiber-optic faceplate, which is tightly connected to a fiber-optic taper. The X-ray-sensitive phosphor surfaces at the front convert the incident X-rays into a burst of visible-light photons. Although it is possible to permit the X-rays to strike the CCD directly, this method has several drawbacks, such as radiation damage to the CCD, signal saturation, and poor efficiency. The use of a larger phosphor as an active detector area and the demagnifying fiber-optic taper is also necessary because the size of the scientific-grade CCD sensors is not as large as needed for the demands of the X-ray diffraction experiment. The fiber-optic taper is then bonded to the CCD, which in turn is connected to the electronic readout system. The CCD must be cooled to temperatures ranging from –40 °C to –90 °C, depending on the various systems. The great advantage of CCD detectors is their short readout time, which lies in the range from 1 to a few seconds.

2.3 Crystal Mounting and Cooling

2.3.1 Conventional Crystal Mounting

The purpose of crystal mounting is to isolate a single crystal from its growth medium so that it can be used in the X-ray diffraction experiment to study its diffraction properties. It is important that the manipulation of the crystal introduces as little damage as possible to its three-dimensional structure. The most important aspect of crystal mounting is to preserve the crystal in its state of hydration. This is accomplished by sealing the crystal in a thin-walled (0.001 mm-thick) glass or quartz capillary tube. The important steps in conventional crystal mounting are illustrated in Figure 2.17 a–c. The crystal must be dislodged from the surface on which it grew, after which it may be drawn into the capillary using suction from a small-volume (0.25 mL) syringe, micropipette or mouth aspirator which are connected to the funnel of the capillary by a flexible plastic hose of appropriate diameter. Next, the capillary should be inverted to allow the crystal to fall to the inner meniscus. The surrounding solution may then be removed using thin strips of filter paper, or with a small glass pipette. The extent to which the crystal should be dried must be determined by experience. The final step is to place a small volume of mother liquor in the capillary and to seal both ends. The capillary is then glued to a metal base which can be attached to a goniometer head.

2.3.2
Cryocrystallography

Many macromolecular crystals suffer from radiation damage when exposed to X-rays with energies and intensities as used in macromolecular X-ray diffraction experiments with both conventional sources and synchrotron radiation. One possibility of reducing radiation damage of the crystal during the measurement is to cool the crystal to low temperatures, usually to 100 K (for a review on cryocrystallography, see Rodgers, 1997). For this purpose, the crystal is flash-frozen to prevent ice formation or damage to the crystal. One method of crystal treatment is to remove the external solution by transferring the crystal in a small drop to a hydrocarbon oil and either teasing the liquid away or drawing it off with filter paper or a small pipette. The oil-coated crystal is then mounted onto a glass-fiber or small glass "spatula". The oil protects the crystal from drying, and also acts as an adhesive that hardens on cooling to hold the sample rigidly in place. A much more frequently used technique is to suspend the crystal in a film of mother liquor in a small loop (Teng, 1990). This method avoids prob-

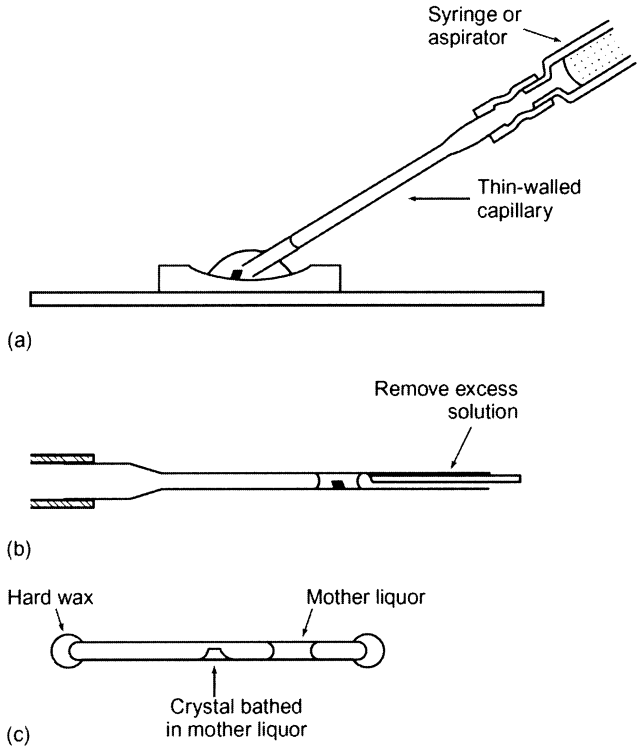


Fig. 2.17 The mounting of a crystal in a glass capillary. (Reproduced by permission of Academic Press, Inc., from Rayment, 1985.)

lems with damage by the oil, or mechanical damage when removing the external liquid, and it has proven successful for most samples. It does, however, require the use of a cryoprotectant to prevent ice formation; the most commonly used cryoprotectants are glycerol, polyethylene glycol (PEG) of different molecular weights, glucose, and 2-methyl-2,4-pentanediol (MPD).

The loop is produced from fine fibers which permit unobstructed data collection in almost all sample orientations. The crystal is held within the loop, suspended in a thin film of cryoprotectant-containing harvest buffer. The loop is supported by a fine wire or pin, which itself is attached to a steel base used for placing the assembly on a goniometer head and in storing mounted crystals. Once in the loop, the crystal is cooled to a temperature at which the increasing

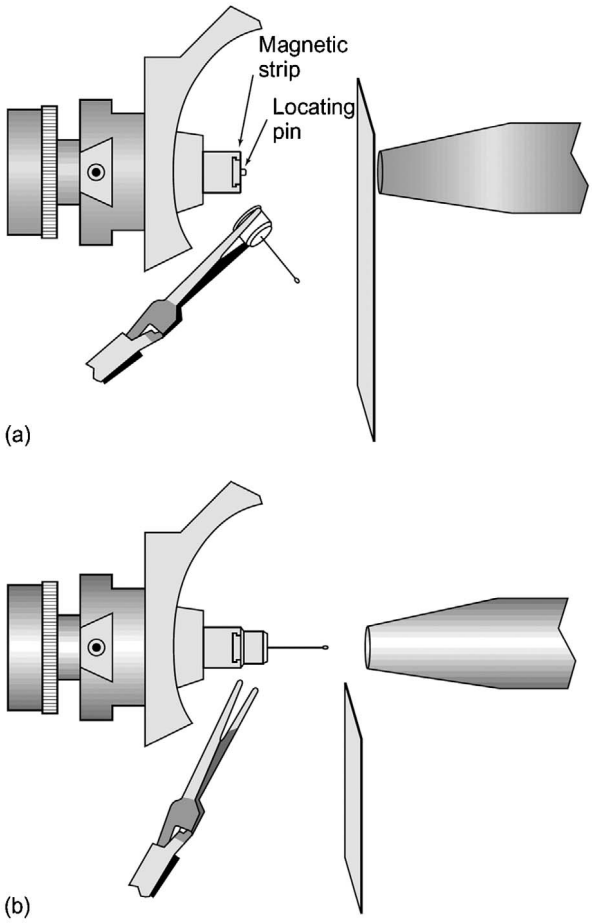


Fig. 2.18 Flash cooling in the direct cold gas stream of a cryostat. (Reproduced by permission of Academic Press, Inc., from Rodgers, 1997.)

viscosity of the liquid prevents molecular rearrangement. The rate of cooling must be rapid enough to reach this point before ice-crystal nucleation occurs. Two methods are used: (i) cooling directly in the gas stream of a cryostat; or (ii) plunging the crystal into a cryogenic liquid. The first method is illustrated in Figure 2.18. The loop assembly (with crystal) is attached to the goniometer head with the cold stream deflected (Fig. 2.18a). The cold stream is then unblocked to flash-freeze the crystal (Fig. 2.18b). In this gas-stream position the goniometer head must be heated to prevent ice formation on the goniometer head. The cryostats and cryocrystallographic tools are all available commercially.

Cryocrystallography has had a major impact on macromolecular crystallography by dramatically increasing the lifetime of a crystal during the X-ray experiment. This allows, for example, the collection of several data sets from one crystal at different wavelengths, using synchrotron radiation. In the case of structural genomics projects, automatic sample changers either for in-house investigations or at protein crystallography (PX) synchrotron beamlines have been developed. An elegant solution for the PX beamlines at ESRF has been developed, and is shown in Figure 2.19. The commercially available vials containing the sample loops with their supports are mounted in baskets, each of which has a capacity for 10 samples. Four of these baskets can be stored in the cryotank of the sample changer. The pincer on the robotic arm moves to the cryotank, takes a vial, and moves it from right to left to the magnetic base of the goniometer head on the goniostat. The support of the loop sticks to the magnetic base of the goniometer head, and the socket of the vial is moved back by the pincer. The loop is immediately placed into the cryo-stream, whereupon the automated centering of the crystal can be started. In the near future, PX beamline users will be able to freeze their samples in cryo loops in-house, ship them in cryo-tanks to the PX beamline, and have the samples tested and measured by web-based remote control from their home laboratory.

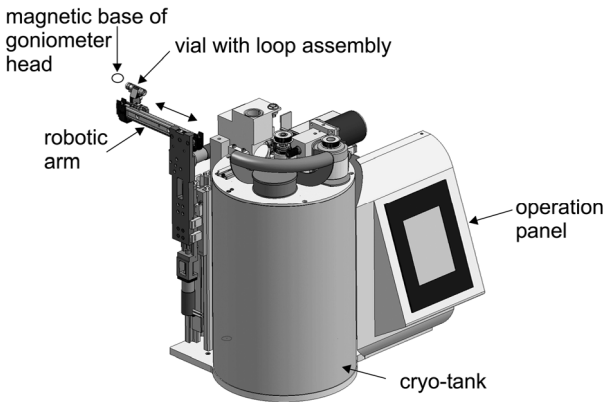


Fig. 2.19 An automatic sample changer used at ESRF PX beamlines, developed by EMBL instrumentation group. (Illustration courtesy of EMBL outstation, Grenoble.)

2.3.3
Crystal Quality Improvement by Humidity Control

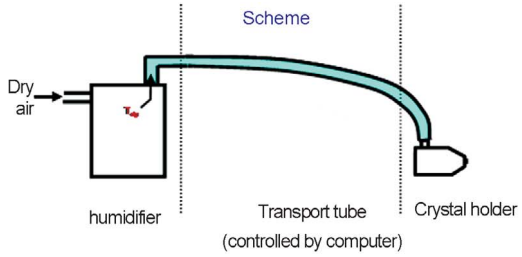
The crystal quality is of decisive significance for a successful X-ray crystal structure determination. Two principal cases must be distinguished: (i) that the crystals diffract to a resolution only (>4.5 Å) which does not allow a structure determination at all; and (ii) that the crystal quality is good enough to elucidate its 3D structure (resolution <3.5 Å, but not better than 2.5 Å). However, a structure determination at higher resolution and accuracy would be necessary, for example in the characterization of a metal center in a metalloprotein. Control of the crystal packing via the solvent content of a crystal is a useful approach to improve the crystal quality of biological macromolecules, and for this purpose a so-called Free Mounting System (FMS) has been developed and successfully applied (Kiefersauer et al., 2000). The principal construction of the FMS is shown in Figure 2.20a. A main part of the FMS is the humidifier unit, which comprises the humidifier, the control, and power electronics. A stream of humid air with defined moisture is produced and transported via a flexible Teflon tube to the crystal holder (Fig. 2.20b). The very compact construction allows the sample to be mounted in a controlled environment with minimal restriction for the X-ray measurement. The head part, into which a heating element and a temperature sensor are integrated, is freely rotatable relative to the insert, without axial movement. The humid air stream through the head part is adjusted to the temperature of the head part independently of the ambient temperature. The crystal may be mounted either in a patch-clamp pipette or a conventional cryo-loop (Fig. 2.20c).

In a typical experiment to increase the crystal quality, X-ray crystal diffraction is monitored at various defined crystal humidities. Usually, a positive effect is observed at lower humidity, which causes shrinkage of the unit cell volume and allows different and possibly more favorable crystal contacts. After having found the optimal condition, the crystal, mounted in a loop, can be shock-frozen for any subsequent data collection. By using the FMS in this way, a remarkable improvement of crystal quality was observed in about 30% of the different projects under investigation (Kiefersauer et al., 2000).

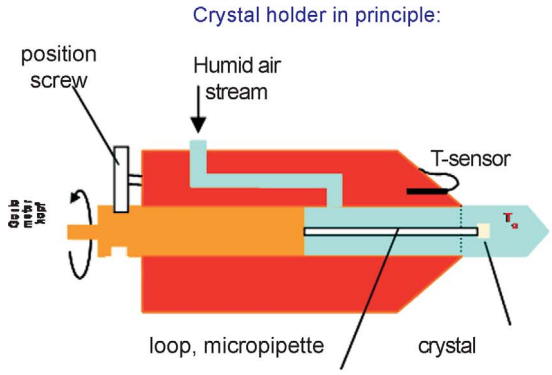
2.4
Data Collection Techniques

2.4.1
Rotation Method

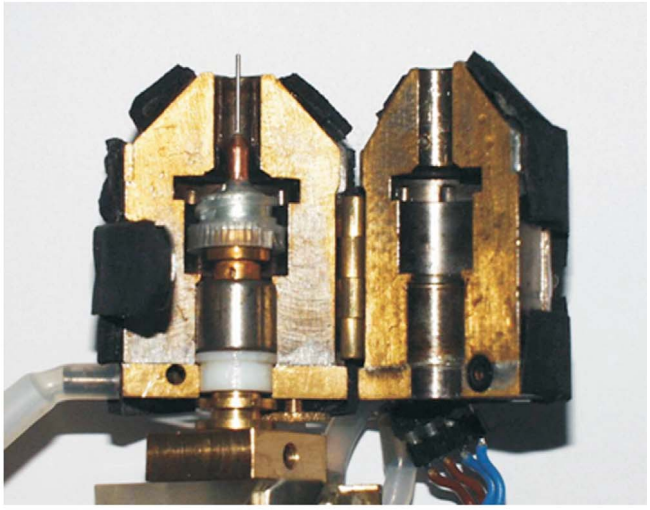
Most macromolecular X-ray diffraction systems use the rotation method for data collection (for a detailed discussion, see Arndt and Wonnacott, 1977). For each crystal, a reciprocal lattice can be constructed which is very useful when interpreting crystallographic crystal diffraction experiments. Diffraction theory (dis-



(a)



(b)



(c)

Fig. 2.20 The Free Mounting System. (a) Principle of construction; (b) schematic view of the crystal holder; (c) the opened crystal holder with magnetic base and mounting loop.

cussed later) tells us that an X-ray reflection is generated when a point of this reciprocal lattice lies on a sphere of radius $1/\lambda$ whose origin is $1/\lambda$ away from the origin of the reciprocal lattice in the direction of the primary beam (Fig. 2.21). The direction of such a diffracted beam is along the connection of the center of the so-called Ewald sphere (radius $1/\lambda$) and the intersection of the reciprocal lattice point on the Ewald sphere. Owing to certain factors (which will be discussed later), the apparent reciprocal lattice extends only to a given radius which defines the resolution sphere. In order to bring all reciprocal lattice points within the resolution sphere into the reflection position, the crystal must be rotated around its center. Almost all macromolecular X-ray diffraction systems apply the rotation technique in the normal beam case, where the rotation axis is normal to the incident X-ray beam. Rotating the crystal around 360° brings all reciprocal lattice points within the resolution sphere in the reflection position, except for the region between the rotation axis and the Ewald sphere; hence, this is called the "blind region". This region can be collected when the crystal has been brought into another orientation. The diffracted beams are usually registered with a flat detector at distance D from the crystal which is also normal to the primary beam. To avoid overlapping of reflection spots on the detector, the crystal is rotated by rotation angle increments; these can vary from tenths of degrees to 1 to a few degrees, depending on the size of the crystal unit cell, crystal mosaicity, beam collimation, and other factors. Each individual exposure is processed and the data stored electronically in a computer. These raw data images are evaluated subsequently with relevant computer pro-

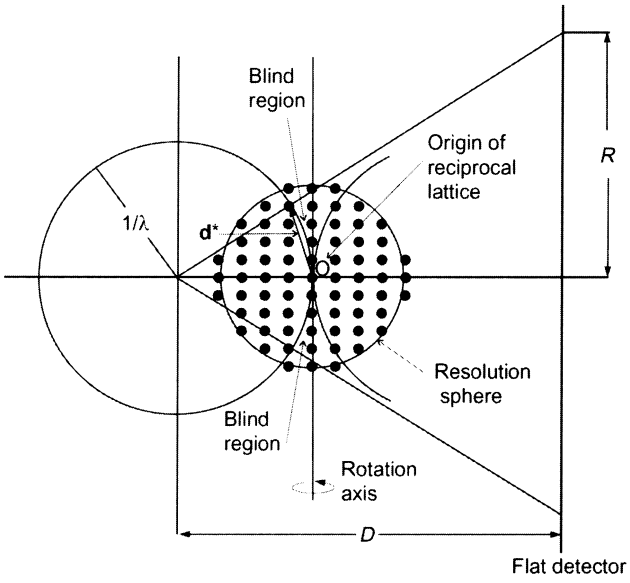


Fig. 2.21 Diffraction geometry in the rotation method usually applied in macromolecular X-ray diffraction systems.

grams (discussed in some detail later) to provide the intensities and geometric reference values (indices) for each collected intensity.

2.4.2
Precession Method

The rotation method delivers a distorted image of the reciprocal lattice for each geometry of the detector (flat or curved) and orientation with respect to the rotation axis. However, an undistorted image of the reciprocal lattice can be obtained by using the precession method. The principle of this technique is shown in Figure 2.22. The detector is a flat film. During the motion of a given reciprocal lattice plane (in Fig. 2.22 a so-called zeroth plane passing through the origin *O* of the reciprocal lattice), the flat detector must always be parallel to this reciprocal lattice plane in order to obtain an undistorted image of this plane. The normal of the reciprocal lattice plane – and consequently also the detector – are inclined with respect to the primary X-ray beam by an angle μ. When the normals of the reciprocal lattice plane and the detector carry out a concerted precession motion of angle μ around the primary X-ray beam, a circular region of the reciprocal lattice plane is registered on the detector (these regions are shown as dashed circles in Fig. 2.22). In a precession camera construction, the crystal and the film cassette are both held in a universal joint; in this way the film and crystal move together in phase with the precession angle, μ. In Figure 2.22 the joints are symbolized as forks, and their linkage by a line. Parallel to the zeroth reciprocal lattice plane is a set of lattice planes that also carry out this precession movement. Those parts of the planes that are swung

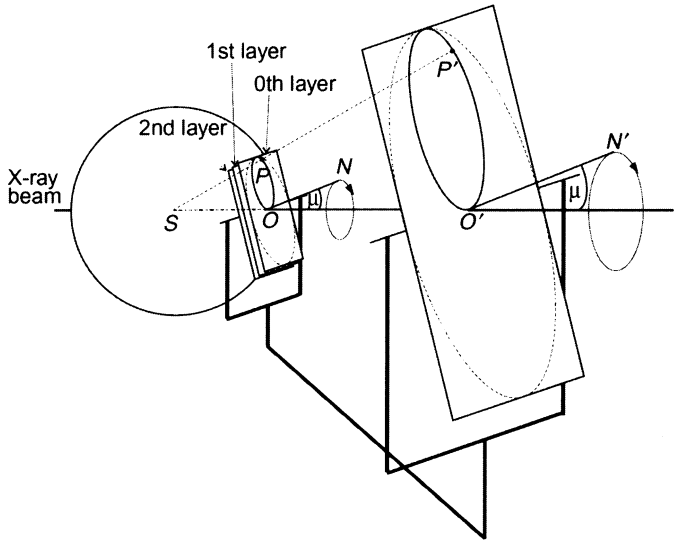


Fig. 2.22 The principle of the precession method (modified from Buerger, 1964).

through the Ewald sphere also give rise to an image on the flat detector (first and second reciprocal lattice layers are also indicated in Fig. 2.22). The images of these layers would superimpose on the detector, and use of the technique in this way is denoted the "screenless precession method". The insertion of a screen with a suitable annular aperture between the crystal and the detector at an appropriate distance can be used to screen out the desired reciprocal layer. This screen is also inclined by the precession angle μ and is coupled to the concerted precession movement of crystal and film cassette. The strength of the precession technique is that, in addition to the undistorted imaging of the reciprocal lattice planes, the indexing of the diffraction spots is straightforward and the symmetry of the diffraction pattern is readily obtained by inspection.

For this reason, the precession method has for a long time been broadly applied in macromolecular crystallography. Although, today, film has been largely replaced by the new generation of detectors, and use of the precession method has almost ceased, the precession camera can still provide students of the subject with a familiarity of the reciprocal lattice concept.

References

Arndt, U.W., Wonnacott, A.J. (Eds.), *The Rotation Method in Crystallography*, North Holland, Amsterdam, **1977**.

Buerger, M.J., *The Precession Method in X-ray Crystallography*, John Wiley & Sons, **1964**.

Buras, B., Tazzari, S. (Eds.), *European Synchrotron Radiation Facility*, ESRP, Grenoble, **1984**.

Helliwell, J.R., *Macromolecular Crystallography with Synchrotron Radiation*, Cambridge University Press, Cambridge, MA, **1992**.

Helliwell, J.R., Synchrotron-radiation instrumentation, methods and scientific utilization. In: M.G. Rossmann, E. Arnold (Eds.), *International Tables for Crystallography*, Vol. F, pp. 155–166. Kluwer Academic Publishers, Dordrecht, **2001**.

Kahn, R., Fourme, R. *Methods Enzymol.* **1997**, *276*, 268–288.

Kiefersauer, R., Than, M.E., Dobbek, H., Gremer, L., Melero, M., Strobl, S., Dias, J.M., Soulimane, T., Huber, R. *J. Appl. Crystallogr.* **2000**, *33*, 1223–1230.

Kirkpatrick, P., Baez, A.V. *J. Opt. Soc. Am.* **1948**, *56*, 1–13.

Moffat, K. *Nature Struct. Biol., Synchrotron Suppl.* **1998**, 641–643.

Phillips, W.C., Rayment, I. *Methods Enzymol.* **1985**, *114*, 316–329.

Rayment, I. *Methods Enzymol.* **1985**, *114*, 136–140.

Rodgers, D.W. *Methods Enzymol.* **1997**, *276*, 183–203.

Teng, T.-Y. *J. Appl. Crystallogr.* **1990**, *23*, 387–391.

Westbrook, E.M., Naday, I. *Methods Enzymol.* **1997**, *276*, 244–268.

3
Principles of X-Ray Diffraction by a Crystal

3.1
Rational Mathematical Representation of Waves

3.1.1
Simple Harmonic Oscillations

The simplest periodic process in time (oscillation) is given mathematically by the sine or cosine function. If the process is recurring ν times per second, it is written as:

$$u(t) = A \sin 2\pi\nu t \quad \text{or} \quad u(t) = A \cos 2\pi\nu t \tag{3.1}$$

where A is the amplitude and ν is the frequency. In many cases, the angular frequency $\omega = 2\pi\nu$ is introduced. The actual value of ωt is called the "phase angle", and this determines the momentary state, the phase. A process described by a simple sine or cosine function is not only mathematically but also physically the simplest oscillation. It is called the harmonic oscillation.

Computation becomes much easier if the imaginary exponential function is used instead of the trigonometric functions, with their cumbersome computing rules. The trigonometric functions are related to the exponential function by *Euler's formula*:

$$\exp 2\pi i\nu t = \cos 2\pi\nu t + i \sin 2\pi\nu t \tag{3.2}$$

This representation leads to a highly important visualization of the oscillation process in an *Argand diagram*, in which

$$z = A \exp 2\pi i\nu t \tag{3.3}$$

a complex number means, the representation point of which rotates on a circle of radius A with an angular velocity of ω. The projections onto the real and imaginary axis are:

$$x = \mathrm{Re}(z) = A \cos 2\pi\nu t; \; y = \mathrm{Im}(z) = A \sin 2\pi\nu t \tag{3.4}$$

As physics is dealing with real magnitudes, the final results of a calculation with complex magnitudes must be translatable into the real. This is done very

easily. An equation between complex numbers means that both the real and the imaginary part of each side fulfill the equation. Thus, we can take both the real and the imaginary parts of the equation as the physical meaning of the equation. Often, the square of the amplitude A^2 is important, and can be obtained in the complex computation most rapidly by multiplying the oscillation magnitude z by its complex conjugated value $\bar{z}$:

$$A^2 = z\bar{z} \tag{3.5}$$

The great advantage of a complex computation emerges if one wants to add two oscillations with the same frequency, but with different phase. Whether there is a single oscillation the start of the calculation of times is irrelevant; therefore, we write it either in the form of Eq. (3.1) or (3.3). If a second oscillation is added to the first one, this oscillation will not reach its maximum at the same moment, but at a certain time before or afterwards. In this case, it can be said that a phase difference δ exists between both oscillations. In the real representation (see Fig. 3.1) we get:

$$u_1(t) = A_1 \cos 2\pi\nu t,\ u_2(t) = A_2 \cos(2\pi\nu t - \delta)$$

and the sum of both oscillations

$$u_1(t) + u_2(t) = A_1 \cos 2\pi\nu t + A_2 \cos\delta \cos 2\pi\nu t + A_2 \sin\delta \sin 2\pi\nu t\,. \tag{3.6}$$

Figure 3.1c shows that not only a new phase φ is created but also that the new amplitude A is not the direct sum of A_1 and A_2.

The last expression can be converted to a single cosine oscillation if one transforms the coefficients of $\sin 2\pi\nu t$ and $\cos 2\pi\nu t$ in such a way that, after separation of a common factor, they gain the property of sine and cosine of the angle φ and that the square sum becomes 1:

$$\begin{aligned}&(A_1 + A_2\cos\delta)\cos 2\pi\nu t + A_2 \sin\delta \sin 2\pi\nu t\\ &\quad = (A_1^2 + A_2^2 + 2A_1A_2\cos\delta)^{1/2}\left\{\frac{A_1 + A_2\cos\delta}{(A_1^2 + A_2^2 + 2A_1A_2\cos\delta)^{1/2}}\cos 2\pi\nu t\right.\\ &\qquad \left. + \frac{A_2\sin\delta}{(A_1^2 + A_2^2 + 2A_1A_2\cos\delta)^{1/2}}\sin 2\pi\nu t\right\}.\end{aligned} \tag{3.7}$$

If one puts

$$\frac{A_1 + A_2\cos\delta}{(A_1^2 + A_2^2 + 2A_1A_2\cos\delta)^{1/2}} = \cos\varphi\,; \tag{3.8}$$

$$\frac{A_2\sin\delta}{(A_1^2 + A_2^2 + 2A_1A_2\cos\delta)^{1/2}} = \sin\varphi\,, \tag{3.9}$$

one gets

$$u_1(t) + u_2(t) = \sqrt{A_1^2 + A_2^2 + 2A_1A_2\cos\delta}\cos(2\pi\nu t - \varphi) = A_{12}\cos(2\pi\nu t - \varphi)\,. \tag{3.10}$$

The formulae can be represented in a vector diagram (Fig. 3.2). The amplitude A_{12} of the resultant oscillation is obtained according to the square-root expres-

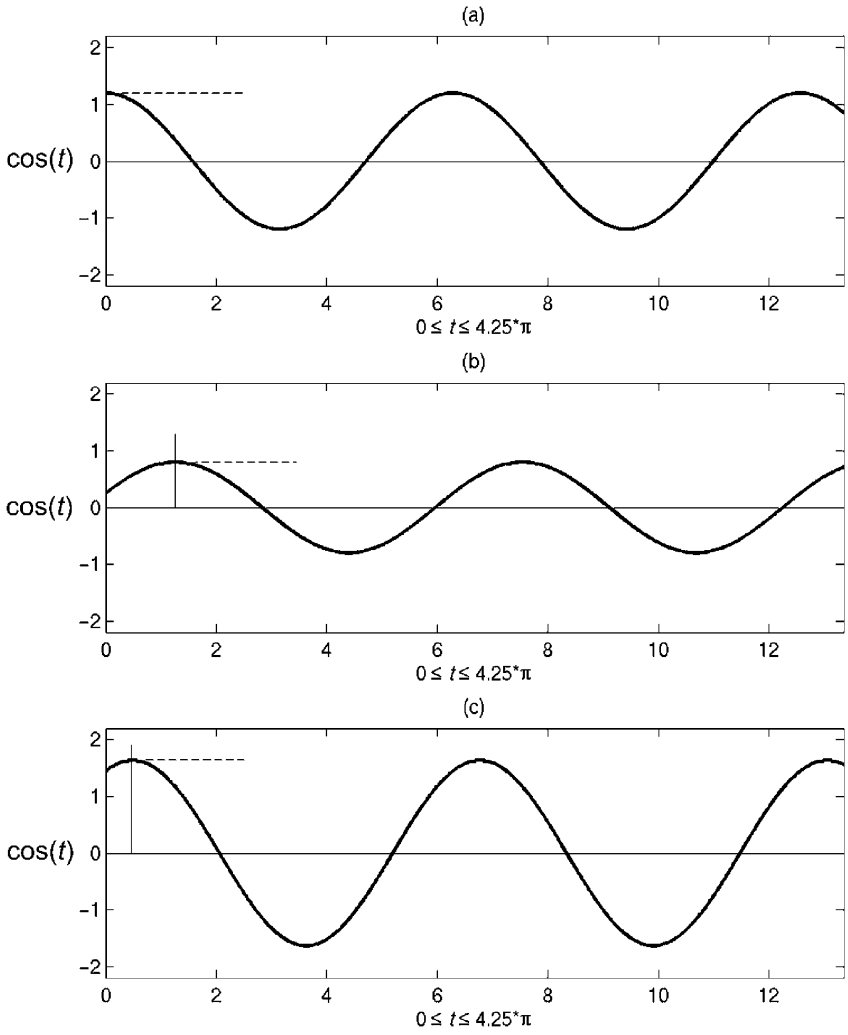


Fig. 3.1 Representation of harmonic oscillations: (a) $u_1(t)$; (b) $u_2(t)$; (c) $u_1(t) + u_2(t)$. (Figure was produced with MATCAB; Math Works, Inc., 2005.)

sion as the third side of a triangle, the other two sides of which are formed by A and B, with angle $(2\pi\nu t - \delta + \varphi)$.

$$\begin{aligned}((A_1 \cos 2\pi\nu t &+ A_2 \cos(2\pi\nu t - \delta))^2 + (A_1 \sin 2\pi\nu t + A_2 \sin(2\pi\nu t - \delta))^2)^{1/2} \\ &= A_1^2(\cos^2 2\pi\nu t + \sin^2 2\pi\nu t) + A_2^2(\cos^2(2\pi\nu t - \delta) + \sin^2(2\pi\nu t - \delta)) \\ &\quad + 2A_1A_2 \cos 2\pi\nu t \cos(2\pi\nu t - \delta) + 2A_1A_2 \sin 2\pi\nu t \sin(2\pi\nu t - \delta) \\ &= (A_1^2 + A_2^2 + 2A_1A_2 \cos\delta)^{1/2}\,, \\ \text{applying } \cos^2\alpha + \sin^2\alpha = 1 &\text{ and } \cos\alpha\cos\beta + \sin\alpha\sin\beta = \cos(\alpha - \beta)\,. \end{aligned} \quad (3.11)$$

The phase difference φ between the resulting oscillation and $u_2(t)$ can also be seen from Figure 3.2. The angle φ is the angle that is formed between A_2 and A_{12}. It can be derived from the triangle formed by A_1, A_2 and A_{12} via the law of cosine.

$$A_1^2 = A_2^2 + A_{12}^2 - 2A_2A_{12}\cos\varphi \quad \cos\varphi = \frac{-A_1^2 + A_2^2 + A_{12}^2}{2A_2A_{12}} \quad (3.12)$$

Substitution of A_{12} and some conversions lead to Eq. (3.8). Using the expression $\sin\varphi = (1 - \cos^2\varphi)^{1/2}$ results in Eq. (3.9). It follows that the sum of two shifted cosine oscillations of equal frequency is again a cosine oscillation with the same frequency, but different phase and amplitude. The used construction completely corresponds to the addition of two complex numbers in the *Argand diagram*,

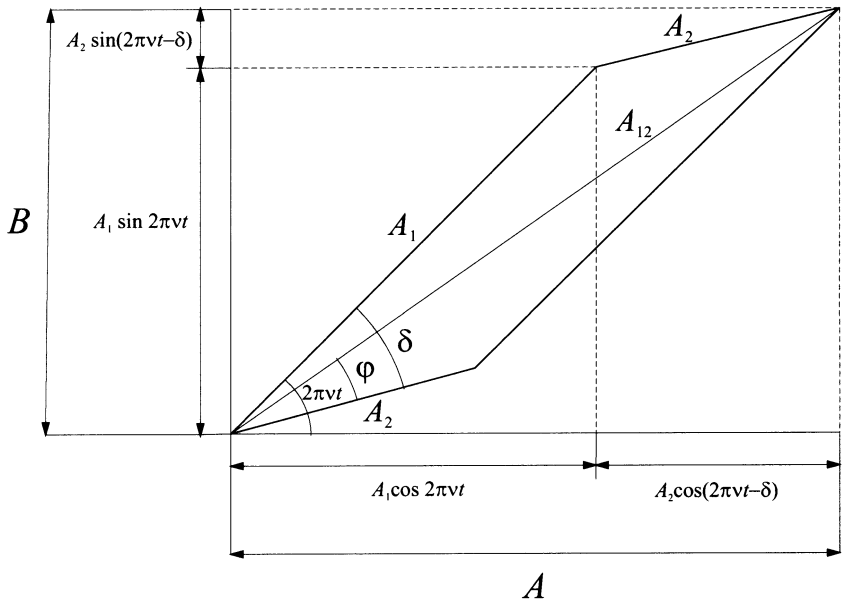


Fig. 3.2 Vector diagram for the addition of $u_1(t) + u_2(t)$.

which are added like vectors, as is generally known. A_1, A_2 and A_{12} in Figure 3.2 can be regarded as pointers rotating on concentric circles centered at the origin of the plot with angular frequency $2\pi\nu$ and constant phase shifts δ and φ, respectively. This construction can be carried out at each moment. The whole parallelogram rotates as a rigid object with angular frequency $2\pi\nu$ because the phase differences are constant. One can choose any position by putting, for example, the first oscillation onto the real axis. The addition of more than two oscillations occurs by vector addition of the individual pointers. The resulting pointer is the connecting passage of the starting point of the first pointer with the end point of the last one. This construction will be used frequently in the respective parts of this book.

3.1.2
Wavelike Propagation of Periodic States

These reflections can now be expanded to all points of space where certain oscillation states may occur. The simplest case is the propagation in a certain direction, taken as *X*-axis, with the state independent of the Y and Z coordinates, which means that the state is a function of x only. At position $x = 0$ we will have

$$u(t, 0) = A \exp 2\pi i\nu t \, . \tag{3.12}$$

We now refer to a process as a wave if the same oscillation state is at position $x = x$ as at position $x = 0$, though with a phase shift which corresponds to the finite propagation velocity v of the phase. The same state at time t is also met in x, which has been met in 0 at time $t - \frac{x}{v}$. Thus it must be written:

$$u(t, x) = A \exp 2\pi i\nu \left(t - \frac{x}{v}\right) \, . \tag{3.13}$$

In the simplest case, the amplitude is independent of the position. Such waves are called "undamped", and only such waves will be considered here. They have constant amplitudes, and u is not only a periodic function of time but also of coordinate x. Imagine a snap-shot of the state at time t_0 for which we get the equation:

$$u(t_0, x) = A \exp 2\pi i\nu t_0 \exp -2\pi i\nu \frac{x}{v} \, . \tag{3.14}$$

The actual value at position 0 recurs if x increases about an integral multiple of $\frac{2\pi v}{2\pi\nu} = \frac{v}{\nu}$, because the argument of the imaginary exponential function is increased by 2π, the functions remains unchanged. The distance between points of equal phase is called the wavelength, λ. According to the reflections above, we obtain the fundamental equation of wave theory:

$$\lambda = \frac{v}{\nu} \quad \text{or} \quad \lambda\nu = v \tag{3.15}$$

wavelength × frequency = velocity of phase propagation

A wave as the actual one, whose variable u depends except for the time of the direction of propagation only, is denoted as a plane wave, as u is constant in a plane perpendicular to the direction of propagation. Figure 3.3 shows the graphical representation of a plane cosine wave $u(t, x)$ with unit amplitude, frequency $\nu = 1\ \mathrm{s}^{-1}$ and $\lambda = 1/2$ length unit propagating in x direction. The time course of one wave state is indicated by solid black lines.

If the normal **n** (i.e., the normal to the planes of equal phase) is arbitrarily oriented to the axes, the planes of equal phase have the equation $\mathbf{rn} = \text{const}$, and the plane wave is represented by

$$u(t, \mathbf{r}) = A \exp 2\pi i \left(t - \frac{\mathbf{rn}}{v}\right) . \tag{3.16}$$

Equation (3.16) is a particulate integral of a most general differential equation, which can be considered as the definition of an undamped wave. This differential equation reads as:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = \Delta u = \frac{1}{v^2} \frac{\partial^2 u}{\partial t^2} \tag{3.17}$$

Another important integral of Eq. (3.17) is the spherical wave, which has a dependency of the distance r from the centre 0 only. This reads:

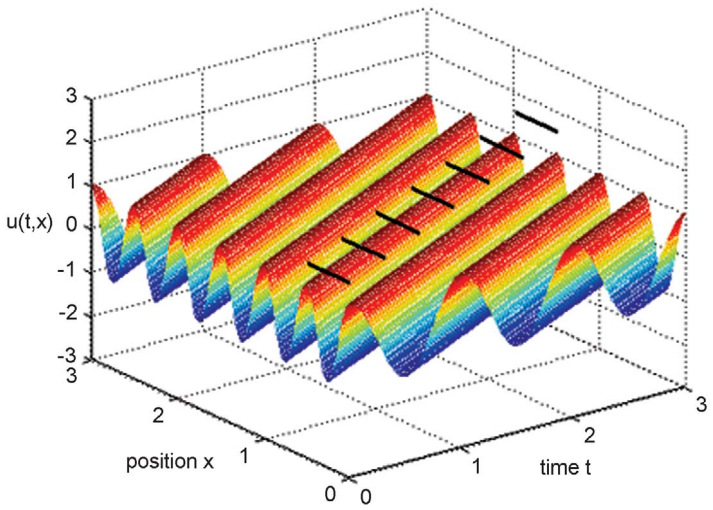


Fig. 3.3 Graphical representation of a plane cosine wave traveling in x direction. (Figure was produced with MATCAB; Math Works, Inc., 2005.)

$$u(t,r) = \frac{A}{r}\exp 2\pi i\left(t - \frac{r}{v}\right) . \tag{3.18}$$

This is a simple harmonic wave, which propagates uniformly to all sides from 0 in the direction of the radius vectors. In contrast to the plane wave, the amplitude decreases with $\frac{1}{r}$ corresponding to the extension of the wave planes. The planes of equal phase are given by $r = \text{const}$; thus, they are spherical shells.

3.2 Principles of X-Ray Diffraction by a Crystal

3.2.1 Scattering of X-Rays by an Electron

In order to obtain a deeper understanding of the physical process of X-ray scattering at atoms, we refer in part to the derivations provided in 1948 by *von Laue*. As X-rays are electromagnetic waves, *Maxwell's equations* are valid here:

$$\begin{aligned} &\text{(A)}\ \operatorname{rot}\mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{H}}{\partial t}, \quad \text{(B)}\ \operatorname{rot}\mathbf{H} = \frac{1}{c}\left(\frac{\partial \mathbf{E}}{\partial t} + 4\pi\mathbf{I}\right) \\ &\text{(C)}\ \operatorname{div}\mathbf{E} = 4\pi\rho , \qquad \text{(D)}\ \operatorname{div}\mathbf{H} = 0 . \end{aligned} \tag{3.19}$$

X-rays propagate in a vacuum with the velocity of light $c = 2.99793 \cdot 10^{10}\ \text{cm s}^{-1}$; they are transversal waves with the electric and magnetic field components **E** and **H**, respectively, oscillating perpendicularly to the direction of propagation and perpendicularly to each other. X-rays cover the spectral range limited by $5 \cdot 10^{-7}\ \text{cm} \geq \lambda \geq 1 \cdot 10^{-11}\ \text{cm}$ or due to $\nu = \frac{c}{\lambda}$ by $6 \cdot 10^{16}\ \text{s}^{-1} \leq \nu \leq 3 \cdot 10^{21}\ \text{s}^{-1}$. With regard to the shortness of their wavelength, we must a priori assume the matter to be atomic complexes rather than continua, and provide the formulae with the adequate shape instantly. Therefore, we speak of the electric vector **E**, the field strength, and the magnetic vector **H**. Only the electric density ρ and the current density **I** occur from the matter here, the latter being the convection current of moved charge carriers (in this case, electrons only). We have then

$$\mathbf{I} = \rho\mathbf{v} \tag{3.20}$$

where **v** is the velocity of the movement. In order to integrate Maxwell's equations it is advisable to introduce a vector potential **A** and a scalar potential Φ according to the approach:

$$\text{(A)}\ \mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{A}}{\partial t} - \operatorname{grad}\Phi \qquad \text{(B)}\ \mathbf{H} = \operatorname{rot}\mathbf{A} . \tag{3.21}$$

This fulfils Eqs. (3.19 A) and (3.19 D) identically. We must also take into account charges and currents in order to make the following approach to fulfill Eqs. (3.19 B) and (3.19 C):

$$\text{(A)}\ \ \Delta\Phi - \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2} = -4\pi\rho, \qquad \text{(B)}\ \ \Delta\mathbf{A} - \frac{1}{c^2}\frac{\partial^2\mathbf{A}}{\partial t^2} = -\frac{4\pi}{c}\mathbf{I}\,,$$
$$\text{(C)}\ \ \operatorname{div}\mathbf{A} + \frac{1}{c}\frac{\partial\Phi}{\partial t} = 0 \tag{3.22}$$

Now, it follows from Eq. (3.21 B):

$$\operatorname{rot}\mathbf{H} = \operatorname{rot}\operatorname{rot}\mathbf{A} = \operatorname{grad}\operatorname{div}\mathbf{A} - \Delta\mathbf{A}$$
$$= -\frac{1}{c}\operatorname{grad}\frac{\partial\Phi}{\partial t} - \frac{1}{c^2}\frac{\partial^2\mathbf{A}}{\partial t^2} + \frac{4\pi}{c}\mathbf{I} = \frac{1}{c}\left(\frac{\partial\mathbf{E}}{\partial t} + 4\pi\mathbf{I}\right)$$

and from Eq. (3.21 A):

$$\operatorname{div}\mathbf{E} = -\frac{1}{c}\frac{\partial}{\partial t}\operatorname{div}\mathbf{A} - \Delta\Phi = \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2} - \Delta\Phi = 4\pi\rho\,,$$

again, in accordance with Eqs. (3.19 B) and (3.19 C). Finally, both potentials can still be traced back to the *Hertz* vector **Z**:

$$\text{(A)}\ \ \Phi = -\operatorname{div}\mathbf{Z}, \qquad \text{(B)}\ \ \mathbf{A} = \frac{1}{c}\frac{\partial\mathbf{Z}}{\partial t}\,, \tag{3.23}$$

because this satisfies Eq. (3.22 C). Equations (3.22 A) and (3.22 B) are satisfied if one sets

$$\Delta\mathbf{Z} - \frac{1}{c^2}\frac{\partial^2\mathbf{Z}}{\partial t^2} = -4\pi\int_0^t \mathbf{I}dt\,. \tag{3.24}$$

As a solution of this wave equation, which corresponds to the waves running out of the limited current area, the following formula can be derived:

$$\mathbf{Z} = \int\frac{d\tau}{r}\int_o^{t-r/c}\mathbf{I}dt\,. \tag{3.25}$$

Here, it must be integrated over the current area, where r represents the distance of the point of integration to the model point. From Eqs. (3.21), (3.23) and (3.24), the equations below follow:

$$\text{(A)}\quad \mathbf{E} = \text{rot rot } \mathbf{Z} - 4\pi \int_0^t \mathbf{I} dt, \qquad \text{(B)}\quad \mathbf{H} = \frac{1}{c} \text{ rot } \frac{\partial \mathbf{Z}}{\partial t} , \tag{3.26}$$

which immediately traces back the field strengths to the *Hertz vector*. By using Eq. (3.24) on most occasions for sine or cosine oscillations, we derive accordingly

$$\mathbf{I} = \mathbf{i}(x, y, z) \exp 2\pi i \nu t , \tag{3.27}$$

Integration for t in this equation delivers a term which is independent of t because the lower integration limit is zero. The independence of t of this term means that it is not related to the oscillation. In contrast, we get

$$\mathbf{Z} = \frac{1}{2\pi i \nu} \int \frac{d\tau}{r} \mathbf{i}(\xi, \eta, \zeta) \exp 2\pi i(\nu t - kr) \qquad \left(k = \frac{\nu}{c}\right) , \tag{3.28}$$

which comes from the upper limit (ξ, η, ζ are the coordinates of the integration point).

We apply this equation to the z-component of the vectors and a model point on the x-axis, the distance of which from the current area is very large compared to the dimensions of the current area. As a consequence, a minimal error is made in the denominator if r is replaced by an average value r_0, though one has to set $r = x - \xi$ in the exponential function. Furthermore, we set the complex quantity

$$\mathbf{i}_z = i_z \exp i\varphi ,$$

with both i_z and φ as spatial function. It follows then:

$$\mathbf{Z}_z(x, 0, 0) = \frac{\exp 2\pi i(\nu t - kx)}{2\pi i \nu r_0} \int d\tau i_z \exp i(2\pi k\xi + \varphi) . \tag{3.29}$$

The *Hertzian solution* of Maxwell's equations follows from Eq. (3.29) by setting $\mathbf{i}$ different to zero in an area, which is small compared to the wavelength $\lambda = 1/k$ and the distance from the model point. If this relates to the periodic oscillation of a point charge $-\varepsilon$ along the z-axis with the amplitude z_0, we get according to Eq. (3.20):

$$\int \mathbf{I}_z = -\varepsilon \frac{dz}{dt} = -2\pi i \nu \varepsilon z_0 \exp 2\pi i \nu t ,$$

thus

$$\int \mathbf{i}_z d\tau = -2\pi i \nu \varepsilon z_0$$

and according to Eq. (3.28):

$$\mathbf{Z}_x = \mathbf{Z}_y = 0;\ \mathbf{Z}_z = -\frac{\varepsilon z_0}{r}\exp 2\pi i(\nu t - kr)\,. \tag{3.30}$$

We are interested in the section of this spherical wave, which propagates in a x'-direction, making the angle δ_z with the z-axis. For that purpose we introduce the coordinates x', $y'=y$ and z', and split the vector $\mathbf{Z}$ into the components

$$\mathbf{Z}_{x'} = \mathbf{Z}_x \cos\delta_z\,,\ \mathbf{Z}_{z'} = \mathbf{Z}_z \sin\delta_z\,.$$

The application of Eqs. (3.26 A) and (3.26 B) would actually require, except for the differentiation of the exponent after the coordinates, also the differentiation of the factor r^{-1}. This delivers higher powers of r^{-1} whereas the other factor $k = \lambda^{-1}$ adds. Thus, one can neglect the differentiation of r^{-1} if the distances are large compared to λ. The wave then behaves like a plane wave, and one finds – as $\mathbf{Z}_{x'}$ is the longitudinal component – that $\mathbf{Z}_{z'}$ is the single transversal component of $\mathbf{Z}$:

$$\mathbf{E}_{z'} = \mathbf{H}_y = (2\pi k)^2 \mathbf{Z}_{z'} = (2\pi k)^2 \mathbf{Z}_z \sin\delta_z,\ \mathbf{E}_y = \mathbf{H}_{z'} = 0 \tag{3.31}$$

Thus, according to Eq. (3.30)

$$\mathbf{E}_{z'} = -(2\pi k)^2 \varepsilon z_0 \frac{\sin\delta_z}{r}\exp 2\pi i(\nu t - kr)\,. \tag{3.32}$$

Now the point charge shall be an electron, which in other respects oscillates freely under the action of the incident wave around its resting position

$$\mathbf{E}_z^e = \mathbf{E}_0^e \exp 2\pi i(\nu t - kx)$$

according to the following equation:

$$\mu\frac{d^2 z}{dt^2} = -\varepsilon \mathbf{E}_z^e\,. \tag{3.33}$$

If x corresponds to the coordinate of this electron, we get

$$z = z_0 \exp 2\pi i\nu t,\ z_0 = \frac{\varepsilon}{(2\pi\nu)^2\mu}\mathbf{E}_0^e \exp -2\pi i k x\,. \tag{3.34}$$

Substituting this into Eq. (3.32), one gets

$$\mathbf{E}_{z'} = -\frac{\varepsilon^2}{c^2\mu}\,\frac{\sin\vartheta_z}{r}\,\mathbf{E}_0^e \exp 2\pi i(\nu t - k(x+r))\,. \tag{3.35}$$

The intensity of the scattered wave propagating along x' is obtained according to Eq. (3.5) with $\mathbf{E}_{z'}$ as complex number, and reveals for the scattering of a wave oscillating in the z-direction *Thomson's scattering formula* for polarized radiation:

$$I_e = \frac{\varepsilon^4}{c^4 \mu^2 r^2} \sin^2 \vartheta_z I_0 \,. \tag{3.36}$$

If an unpolarized radiation of intensity I_0 hits the electron, the wave oscillating in the z-direction has an intensity $\frac{1}{2} I_0$ only. In exchange, a scattering intensity of an oscillation in the y-direction is added with an inclination ϑ_y of the scattering direction against the y-axis. Consequently, we obtain Thompson's scattering formula for unpolarized radiation:

$$I_e = \frac{\varepsilon^4}{c^4 \mu^2 r^2} \frac{\sin^2 \vartheta_y + \sin^2 \vartheta_z}{2} I_0 = \frac{\varepsilon^4}{c^4 \mu^2 r^2} \frac{1 + \cos^2 \vartheta_x}{2} I_0 \,. \tag{3.37}$$

The transformation is based on the identity:

$$1 = \cos^2 \vartheta_x + \cos^2 \vartheta_y + \cos^2 \vartheta_z = \cos^2 \vartheta_x + 2 - \sin^2 \vartheta_y - \sin^2 \vartheta_z \,.$$

The direction factor $1/2(1 + \cos^2 \vartheta)$ is denoted as the "polarization factor", where ϑ is the scattering angle between the direction of the incident beam and the scattering direction. Frequency ν no longer occurs in Eqs. (3.26) and (3.27).

3.2.2 Scattering of X-Rays by an Atom

In this section, it is the electrons of an atom that cause the scattering. The electrons are numbered by the index n, where the n-th electron may have the coordinate $x_0 + x_n$ with x_0 the coordinate of an arbitrary reference point O in the atom (Fig. 3.4a). The field strength, which is generated by the scattering at the n-th atom as according to Eq. (3.35):

$$\mathbf{E}_{z'} = -\frac{\varepsilon^2}{c^2 \mu} \frac{\sin \vartheta_z}{r_n} \mathbf{E}_0^e \exp 2\pi i(\nu t - k(r_n + x_0 + x_n)) \,. \tag{3.38}$$

We introduce some useful vectors to avoid any special choice of coordinate system. We define the wave vector $\mathbf{s}_0$, which is parallel to the direction of the incident beam and has an absolute value of $s_0 = 1/\lambda = k$. Figure 3.4b shows the situation with the x, y, z-coordinate system in an arbitrary position and $\mathbf{s}_0$ parallel to the x-direction in Figure 3.4a. We can substitute kx_0 in Eq. (3.38) by $\mathbf{s}_0 \mathbf{r}_{0'}$ because $x_0 = r_{0'} \cos(\mathbf{s}_0, \mathbf{r}_{0'})$ and $kx_0 = s_0 x_0 = s_0 r_{0'} \cos(\mathbf{s}_0, \mathbf{r}_{0'})$.

We determine the position of the n-th electron by the radius vector $\mathbf{R}_n$ traced from the reference point O in the atom (Fig. 3.5). The observation point on the detector is linked with to reference point O in the atom by vector $\mathbf{r}_0$, and the scattered wave generated by the oscillation electron E_n by vector $\mathbf{r}_n$. The wave vector $\mathbf{s}$ is parallel to $\mathbf{r}_0$ and has the same absolute value as $\mathbf{s}_0$ (Fig. 3.5). We substitute kr_n and kx_n by the scalar products of the relevant vectors as we did it for kx_0 and obtain for the scattered wave of the n-th electron:

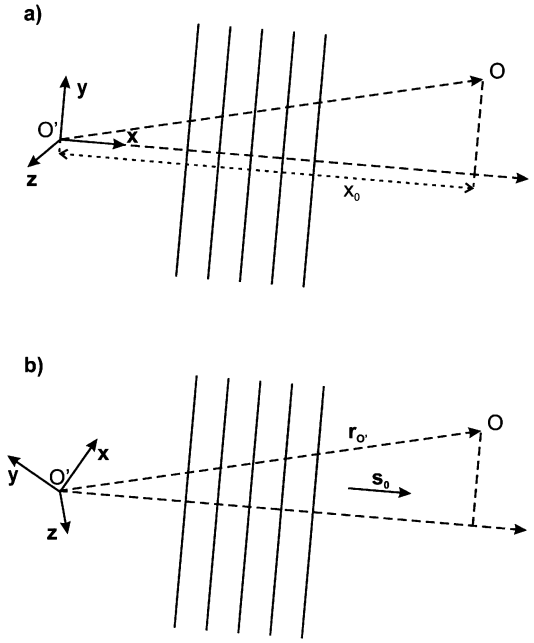


Fig. 3.4 (a) Origin O of an atom in coordinate system *x*, *y*, *z* with a planar X-ray wave traveling in *x* direction. (b) The same situation as in (a), but the *x*, *y*, *z* coordinate system is in an arbitrary position and the plane wave is traveling in a direction determined by wave vector $\mathbf{s}_0$, which is parallel to the *x* direction in (a).

$$\mathbf{E}^n_{z'} = -\frac{\varepsilon^2}{c^2\mu}\,\frac{\sin\vartheta_z}{r_n}\,\mathbf{E}^e_0 \exp 2\pi i(\nu t - \mathbf{s}_0\mathbf{r}_{0'} - \mathbf{s}_0\mathbf{r}_n - \mathbf{s}_0\mathbf{R}_n)\ . \tag{3.39}$$

Equations (3.40) to (3.43) follow from Figure 3.5:

$$\mathbf{R}_n + \mathbf{r}_n = \mathbf{r}_0 \tag{3.40}$$

$$r_n^2 = (\mathbf{r}_0 - \mathbf{R}_n)^2 = r_0^2 + R_n^2 - 2r_0R_n\cos(\mathbf{R}_n, \mathbf{r}_0) \tag{3.41}$$

$$r_n = r_0\left[1 - \frac{2R_n}{r_0}\cos(\mathbf{R}_n, \mathbf{r}_0) + \left(\frac{R_n}{r_0}\right)^2\right]^{1/2} \tag{3.42}$$

$$r_n \approx r_0\left(1 - \frac{R_n}{r_0}\cos(\mathbf{R}_n, \mathbf{r}_0) + \ldots\right) \tag{3.43}$$

where $(R_n/r_0)^2$ and higher terms were neglected in the expansion of the square root. It follows that (Eq. 3.44)

$$r_n \approx r_0 - R_n\cos(\mathbf{R}_n, \mathbf{r}_0)\ . \tag{3.44}$$

We can replace r_n in the exponent by this expression. As $s_0 = s$ and $\mathbf{s}$ is parallel to $\mathbf{r}_0$, we obtain Eq. (3.45).

$$s_0 R_n \cos(\mathbf{R}_n, \mathbf{r}_0) = sR_n \cos(\mathbf{R}_n, \mathbf{s}) = \mathbf{s}\mathbf{R}_n \tag{3.45}$$

As r_0 is large compared to R_n, r_n can be replaced by r_0 in the denominator of Eq. (3.39). With these approximations we obtain from Eq. (3.39):

$$\mathbf{E}^n_{z'} = -\frac{\varepsilon^2}{c^2\mu}\,\frac{\sin\vartheta_z}{r_n}\,\mathbf{E}^e_0 \exp 2\pi i(\nu t - \mathbf{s}_0\mathbf{r}_{0'} - \mathbf{s}_0\mathbf{r}_0 + \mathbf{s}\mathbf{R}_n - \mathbf{s}_0\mathbf{R}_n)\ . \tag{3.46}$$

If we carry out the summation over all electrons, we obtain for the scattered wave from the whole atom:

$$\mathbf{E}_{z'} = \sum_n \mathbf{E}^n_{z'} = -\frac{\varepsilon^2}{c^2\mu}\,\frac{\sin\vartheta_z}{r_0}\,\mathbf{E}^e_0 \exp 2\pi i(\nu t - \mathbf{s}_0(\mathbf{r}_{0'} + \mathbf{r}_0)) \sum_n \exp 2\pi i(\mathbf{R}_n, \mathbf{s} - \mathbf{s}_0)\ . \tag{3.47}$$

The resultant of the sum in Eq. (3.47) is a complex number with an absolute value C, which is in general unequal to 1 and characteristic for the scattering from the atom for the wave incident in direction $\mathbf{s}_0$ and scattered in direction $\mathbf{s}$. Equation (3.47) represents the multiplication of two complex numbers with individual absolute values. The physical meaning is that a complex number with an absolute value of $\frac{\varepsilon^2}{c^2\mu}\,\frac{\sin\vartheta_z}{r_0}\,\mathbf{E}^e_0 \cdot C$ rotates in time t with frequency ν in the *Argand diagram*, which represents the electric field component $\mathbf{E}_{z'}$. As the first expression is constant for the scattering experiment, the individual values for C determine the physical real magnitudes such as the electric field strength $|\mathbf{E}_{z'}|$ or the intensity $|\mathbf{E}_{z'}|^2$. If the scattering of a volume element dv is proportional to the local electron density $\rho(\mathbf{r})$, then the scattering amplitude will be proportional to the integral:

$$f(\mathbf{S}) = \int_{\text{vol. of atom}} \rho(\mathbf{r}) \exp(2\pi i \mathbf{r}\mathbf{S})\, dv \tag{3.48}$$

with $\mathbf{S} = \mathbf{s} - \mathbf{s}_0$ and $\mathbf{r}$ replacing the individual positional vectors $\mathbf{R}_n$.

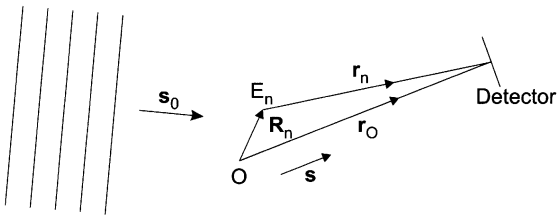


Fig. 3.5 Scattering of a planar X-ray wave traveling parallel to $\mathbf{s}_0$ by an electron E_n at position $\mathbf{R}_n$.

3.2.3
The Atomic Scattering Factor

Equation (3.48) is the general form of the elastic scattering of X-rays by an atom, and is called the "atomic scattering factor". Now, we assume that $\rho(\mathbf{r})$ depends only on the magnitude of **r**, and not on its direction, so that the electron density distribution within the atom is spherically symmetrical. For this case the execution of the integration illustrated in Figure 3.6 is performed in the following manner. The surface element is $rdadr$; this is traced along the circumference $2\pi r \sin a$, and consequently we obtain for the volume element $d\tau = 2\pi r^2 \sin a da dr$. With

$$2\pi \mathbf{rS} = 4\pi \frac{\sin\theta}{\lambda} r \cdot \cos a = \mu r \cos a = x, \left(\mu = 4\pi \frac{\sin\theta}{\lambda}\right), \tag{3.49}$$

we can write:

$$f = \int \rho(r) \cdot \exp ix \cdot d\tau \,. \tag{3.50}$$

Now, we introduce x in $d\tau$:

$$x = \mu r \cos a,\ dx = -\mu r \sin a da,\ d\tau = 2\pi r^2 dr - \left(\frac{dx}{\mu r}\right). \tag{3.51}$$

Thus, we get:

$$f = \int \frac{2\pi r^2}{\mu r} \rho(r) dr \exp ix \cdot (-dx) \,. \tag{3.52}$$

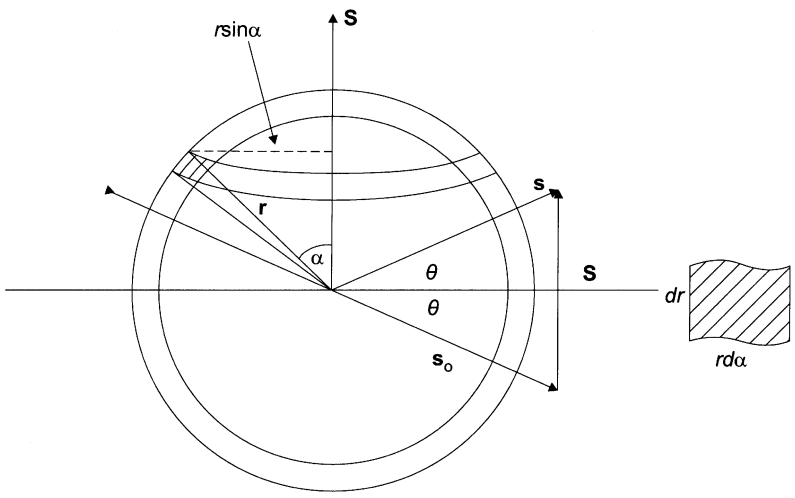


Fig. 3.6 A schematic diagram to explain the integration of Eq. (3.48).

If a runs from 0 to π then x from $+\mu r$ to $-\mu r$. The integration limits are now due to $(-dx)$

$$f = \int_0^{\infty} \frac{2\pi r^2}{\mu r} \rho(r) \cdot dr \int_{-\mu r}^{+\mu r} \exp ix \cdot dx \tag{3.53}$$

The integral reveals $\int_{-\mu r}^{+\mu r} \exp ix \cdot dx = \frac{\exp i\mu r - \exp -i\mu r}{i} = 2 \sin \mu r$ and we get for f:

$$f = \int_0^{\infty} U(r) \frac{\sin \mu r}{\mu r} dr \quad \text{with} \quad U(r) = 4\pi r^2 \rho(r) \tag{3.54}$$

The functions $U(r)$ are the radial densities of the electrons of an atom. Function $\frac{\sin \mu r}{\mu r}$ with $\mu = \frac{4\pi}{\lambda} \sin \theta$ is 1 for all r at $\frac{\sin \theta}{\lambda} = 0$. It is null if $\mu r = \pi, 2\pi, \ldots$.

The radial densities of electrons are calculated by quantum mechanical methods, either by the self-consistent field method or according to the statistical method of Thomas and Fermi. A diagram of the atomic scattering factor of various chemical elements is shown in Figure 3.7. The scattering power is equal to

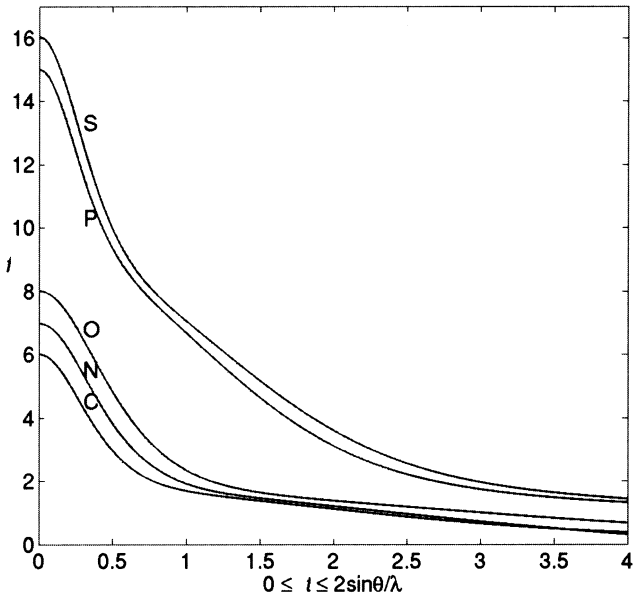


Fig. 3.7 Atomic scattering factors for main non-hydrogen elements contained in biomacromolecules. (Figure was produced with MATLAB; MathWorks, Inc., 2005.)

the number Z of electrons of the element at $\theta = 0$, and decreases with increasing values of θ dependent on $\frac{\sin\theta}{\lambda}$. Tables of the atomic scattering factors of all chemical elements can be found in Volume C of the International Tables for Crystallography (Prince, 2004). These have been incorporated into the relevant crystallographic computer programs.

3.2.4 Scattering of X-Rays by a Unit Cell

A unit cell may contain N atoms at positions of their internal origins at $\mathbf{r}_j$ ($j = 1, 2, 3, \ldots, N$) with respect to the origin of the unit cell (Fig. 3.8). For atom 1, we obtain:

$$\mathbf{f}_1 = \int_{\text{vol. of atom}} \rho(\mathbf{r}) \exp[2\pi i(\mathbf{r}_1 + \mathbf{r})\mathbf{S}]\mathrm{d}v = f_1 \exp(2\pi i \mathbf{r}_1 \mathbf{S}) \tag{3.55}$$

with (Eq. 3.56)

$$f_1 = \int_{\text{vol. of atom}} \rho(\mathbf{r}) \exp 2i\mathbf{r}\mathbf{S}\mathrm{d}v \tag{3.56}$$

where f_1 is the atomic scattering factor for atom 1. This reflects the characteristics of the scattering of the individual atoms, and is real if the wavelength of the incident X-ray is not close to an absorption edge of the atom.

For N atoms this adds up to the total scattered wave of a unit cell $\mathbf{F}(\mathbf{S})$ (Fig. 3.9) according to Eq. (3.57):

$$\mathbf{F}(\mathbf{S}) = \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{r}_j \mathbf{S}) \tag{3.57}$$

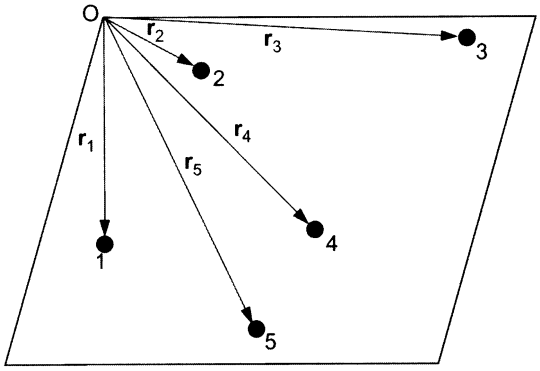


Fig. 3.8 Atomic positions in a unit cell.

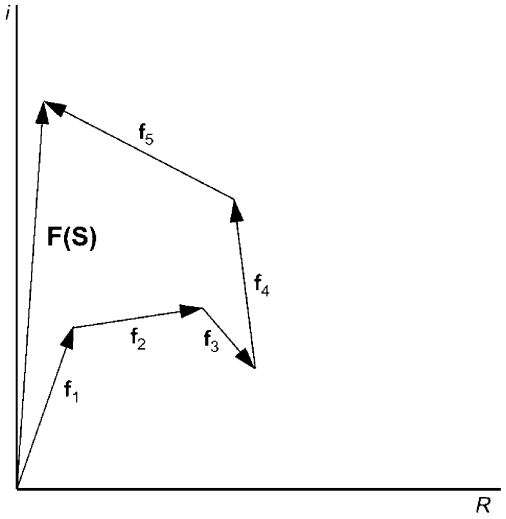


Fig. 3.9 Vector diagram for the total scattered wave in direction **S** added up for five atoms.

3.2.5 Scattering of X-Rays by a Crystal

3.2.5.1 One-Dimensional Crystal

In a one-dimensional crystal the unit cells are separated by the unit cell vector **a**. The contribution of the scattered wave from the unit cell at the origin of the crystal is **F**(**S**). All scatterers in the second unit cell are displaced by the vector **a** relative to the origin, which introduces a corresponding phase factor and reveals for the second unit cell relative to the origin $\mathbf{F}(\mathbf{S})\exp 2\pi i\mathbf{aS}$.

For the nth unit cell relative to the origin we obtain $\mathbf{F}(\mathbf{S})\exp 2\pi i\,(n-1)\,\mathbf{aS}$.

This sums up for the total wave to Eq. (3.58):

$$\mathbf{E}(\mathbf{S}) = \sum_{n=1}^{T} \mathbf{F}(\mathbf{S}) \exp 2\pi i(n-1)\mathbf{aS} \tag{3.58}$$

Generally, **E**(**S**) is of the same order of magnitude as **F**(**S**), and no strong scattering effect is observed (Fig. 3.10 a). However, when $2\pi\mathbf{aS} = 2\pi h$ or an integral multiple of 2π or $\mathbf{aS} = h$ (h is an integer), the waves add up constructively to a scattered wave proportional to $T|\mathbf{F}(\mathbf{S})|$ (Fig. 3.10 b).

3.2.5.2 Three-Dimensional Crystal

In this case, the unit cell is spanned by the unit cell vectors **a**, **b** and **c** and is repeated periodically by the corresponding vector shifts $\mathbf{r} = m_1\mathbf{a} + m_2\mathbf{b}$

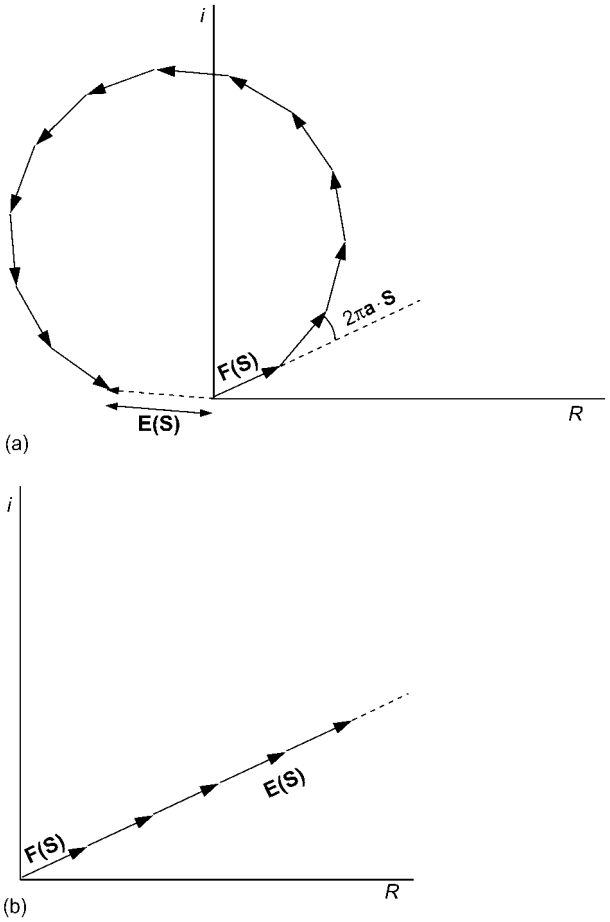


Fig. 3.10 Vector diagrams displaying the total wave scattered by a molecule in a crystal. (a) The phase differences between waves scattered by adjacent unit cells is $2\pi\mathbf{aS}$. (b) The phase difference is an integral multiple of 2π. (Adapted from Blundell and Johnson, 1976.)

$+m_3\mathbf{c}$ (m_1, m_2, m_3, integers) in the respective spatial directions. We can then write for $\mathbf{E(S)}$

$$\mathbf{E}(\mathbf{S}) = -\frac{\varepsilon^2}{c^2\mu}\,\frac{1}{r_0}\sqrt{\frac{1+\cos^2 2\theta}{2}}\,\mathbf{E}_0^e\mathbf{F}(\mathbf{S})\sum_{m_1}^{N_1}\exp 2\pi i m_1(\mathbf{aS}) \cdot \sum_{m_2}^{N_2}\exp 2\pi i m_2(\mathbf{bS})\sum_{m_3}^{N_3}\exp 2\pi i m_3(\mathbf{cS}) \tag{3.58.1}$$

The three sum terms are geometric series which can be easily calculated. For the first sum we get:

$$\sum_{m_1=0}^{N_1} \exp 2\pi i m_1(\mathbf{aS}) = \frac{\exp 2\pi i N_1(\mathbf{aS}) - 1}{\exp 2\pi i(\mathbf{aS}) - 1} \tag{3.58.2}$$

The other sums are calculated analogously. We obtain the intensity or flux by squaring **E(S)**. The squares of the complex magnitudes are received by multiplying it with its complex conjugate. Carrying out these multiplications reveals concise expressions and we get for the intensity $I(\mathbf{S})$:

$$I(\mathbf{S}) = \frac{\varepsilon^4}{c^4\mu^2}\,\frac{1}{r_0^2}\left(\frac{1+\cos^2\theta}{2}\right) I_0 F^2(\mathbf{S})\,\frac{\sin^2\pi N_1(\mathbf{aS})}{\sin^2\pi(\mathbf{aS})}\,\frac{\sin^2\pi N_2(\mathbf{bS})}{\sin^2\pi(\mathbf{bS})}\frac{\sin^2\pi N_3(\mathbf{cS})}{\sin^2\pi(\mathbf{cS})} \tag{3.58.3}$$

The last three factors are known as interference function I_F, which has interesting properties. The function has maxima of $N_1^2 N_2^2 N_3^2$ when the three subsequent conditions are fulfilled (Eq. 3.59):

$$\mathbf{aS} = h;\ \mathbf{bS} = k;\ \mathbf{cS} = l \tag{3.59}$$

These conditions are known as *Laue equations*. The first term becomes, for example, $\frac{\sin^2\pi N_1 h}{\sin^2\pi h}$. The maximum of this function is obtained by twofold subsequent differentiation for h, and reveals in this case N_1^2. The function has subsidiary maxima for non-integral values of h, k, l with zeros between the maxima. The larger the number of unit cells in each crystal dimension N_1, N_2, N_3 is, the closer are the subsidiary maxima to the main maximum and decrease to values close to zero very rapidly. This means that there are sharp intensity maxima at or in close vicinity to the integrals h, k, l, and negligible intensity between them.

If we neglect the constant magnitudes in Eq. (3.58.1), we obtain Eq. (3.60) for the total scattered wave for a three-dimensional crystal with a unit cell containing N atoms:

$$\mathbf{F}(\mathbf{S}) = \sum_{j=1}^{N} f_j \exp 2\pi i \mathbf{r}_j\mathbf{S} \tag{3.60}$$

with (Eq. 3.61)

$$\mathbf{r}_j = \mathbf{a}x_j + \mathbf{b}y_j + \mathbf{c}z_j \tag{3.61}$$

Hence we have Eqs. (3.62) and (3.63):

$$\mathbf{r}_j\mathbf{S} = x_j\mathbf{aS} + y_j\mathbf{bS} + z_j\mathbf{cS} = hx_j + ky_j + lz_j \tag{3.62}$$

(from Laue's equation) and

$$\mathbf{F}(hkl) = \sum_{j=1}^{N} f_j \exp 2\pi i(hx_j + hy_j + lz_j) = |\mathbf{F}(hkl)| \exp i\alpha(hkl) \tag{3.63}$$

with $|\mathbf{F}(hkl)|$ – amplitude and α – phase angle. We obtain the intensity of the scattered wave as the structure factor $\mathbf{F}(hkl)$ multiplied by its complex conjugate value according to (Eq. (3.64)):

$$I(hkl) = \mathbf{F}(hkl)\mathbf{F}^*(hkl) = |\mathbf{F}(hkl)|^2 . \tag{3.64}$$

3.2.6
The Reciprocal Lattice and Ewald Construction

The usefulness of the concept of the reciprocal lattice in understanding the diffraction of X-rays from a crystal was outlined in Section 2.4. Now we have the necessary relationships to derive the reciprocal lattice. One can write the scattering vector **S** as:

$$\mathbf{S} = h_x\mathbf{a}^* + k_y\mathbf{b}^* + l_z\mathbf{c}^* \tag{3.65}$$

where **S** is a vector in reciprocal space with the metric $\mathbf{a}^*$, $\mathbf{b}^*$, and $\mathbf{c}^*$. The relationship to the direct space with metric **a**, **b**, and **c** is still unknown. The vector **S** must obey the Laue equations:

$$\mathrm{a}\mathbf{S} = \mathbf{a}(h_x\mathbf{a}^* + k_y\mathbf{b}^* + l_z\mathbf{c}^*) = h = h_x\mathbf{a}\mathbf{a}^* + k_y\mathbf{a}\mathbf{b}^* + l_z\mathbf{a}\mathbf{c}^* = h \tag{3.66}$$

This is fulfilled only when $\mathbf{a}\mathbf{a}^* = 1$, $h_x = h$ and $\mathbf{a}\mathbf{b}^*$ and $\mathbf{a}\mathbf{c}^* = 0$. Similar equations can be derived for the other two Laue conditions. Thus, vector **S** is a vector of a lattice in reciprocal space. The relationship between the direct and reciprocal lattices is given by the following set of nine equations (Eqs. 3.67):

$$\begin{aligned}
&\mathbf{a}\mathbf{a}^* = 1 \qquad \mathbf{b}\mathbf{a}^* = 0 \qquad \mathbf{c}\mathbf{a}^* = 0\\
&\mathbf{a}\mathbf{b}^* = 0 \qquad \mathbf{b}\mathbf{b}^* = 1 \qquad \mathbf{c}\mathbf{b}^* = 0\\
&\mathbf{a}\mathbf{c}^* = 0 \qquad \mathbf{b}\mathbf{c}^* = 0 \qquad \mathbf{c}\mathbf{c}^* = 1
\end{aligned} \tag{3.67}$$

It follows from these that $\mathbf{a}^* \perp \mathbf{b}$; $\mathbf{c}$; $\mathbf{b}^* \perp \mathbf{a}$; $\mathbf{c}$; $\mathbf{c}^* \perp \mathbf{a}$; $\mathbf{b}$; and vice versa. The metric relationships can also be derived from these relationships. They adopt the following form for the general case of the triclinic crystal system:

$$V = \frac{1}{V^*} = abc\sqrt{1 - \cos^2\alpha - \cos^2\beta - \cos^2\gamma + 2\cos\alpha\cos\beta\cos\gamma}$$

$$V^* = \frac{1}{V} = a^* b^* c^* \sqrt{1 - \cos^2\alpha^* - \cos^2\beta^* - \cos^2\gamma + 2\cos\alpha^*\cos\beta^*\cos\gamma^*}$$

$$a^* = \frac{bc\sin\alpha}{V} \qquad a = \frac{b^* c^* \sin\alpha^*}{V^*}$$

$$b^* = \frac{ac\sin\beta}{V} \qquad b = \frac{a^* c^* \sin\beta^*}{V^*}$$

$$c^* = \frac{ab\sin\gamma}{V} \qquad c = \frac{a^* b^* \sin\gamma^*}{V^*}$$

$$\cos\alpha^* = \frac{\cos\beta\cos\gamma - \cos\alpha}{\sin\beta\sin\gamma} \qquad \cos\alpha = \frac{\cos\beta^*\cos\gamma^* - \cos\alpha^*}{\sin\beta^*\sin\gamma^*}$$

$$\cos\beta^* = \frac{\cos\alpha\cos\gamma - \cos\beta}{\sin\alpha\sin\gamma} \qquad \cos\beta = \frac{\cos\alpha^*\cos\gamma^* - \cos\beta^*}{\sin\alpha^*\sin\gamma^*}$$

$$\cos\gamma^* = \frac{\cos\alpha\cos\beta - \cos\gamma}{\sin\alpha\sin\beta} \qquad \cos\gamma = \frac{\cos\alpha^*\cos\beta^* - \cos\gamma^*}{\sin\alpha^*\sin\beta^*} \tag{3.68}$$

This means that the inverse lattice vectors are perpendicular to the plane, which is spanned by the two other non-inverse lattice vectors. Bragg's law can now be derived by inspection of Figure 3.11. The wave vectors for the incident wave $\mathbf{s}_0$ and the scattered wave s have the same absolute value of $1/\lambda$. Vector **S** must be a vector of the reciprocal lattice, and its absolute value is equal to d^*. From Figure 3.11 we obtain Eqs. (3.69–3.71):

$$\sin\theta = \frac{d^*}{2}\lambda \tag{3.69}$$

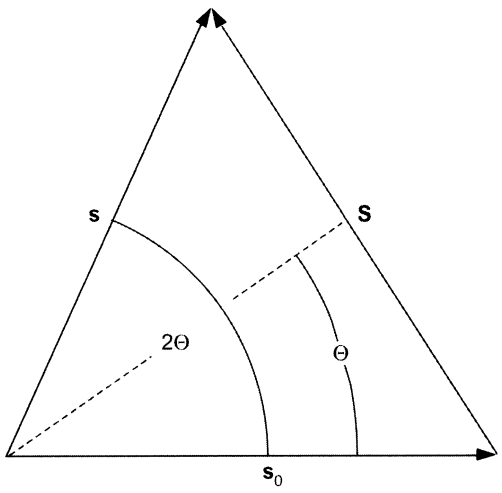


Fig. 3.11 Geometric representation of diffraction geometry, 2Θ, glance angle; Θ, Bragg angle.

$$\lambda = \frac{2\sin\theta}{d^*} \tag{3.70}$$

$$\lambda = 2d\sin\theta \quad \text{for} \quad n = 1 \tag{3.71}$$

The general equation for Bragg's law is:

$$2d\sin\theta = n\lambda \tag{3.72}$$

where n is the order of reflection and d the interplanar distance in the direct lattice.

The Ewald construction is contained in Figure 2.21. A sphere of radius $1/\lambda$ is drawn, and the origin of the reciprocal lattice is located where the wave vector $\mathbf{s}_0$ ends on the Ewald sphere. A diffracted beam is generated if a reciprocal lattice vector $\mathbf{d}^*_{hkl}$ with an absolute value of $1/d_{hkl}$ cuts the Ewald sphere. The beam is diffracted in the direction of the connection of the origin of the Ewald sphere and the intersection point of the reciprocal lattice point on the Ewald sphere. The diffraction pattern of a lattice is itself a lattice with reciprocal lattice dimensions.

3.2.7
The Temperature Factor

At first glance it might appear that the thermal motions of the atoms destroy the sharp reflections deduced for the scattering of X-rays by a crystal. However, as shown by Debye, one observes sharp reflections further on, but the thermal motion of the atoms causes a decrease in the scattering power. To derive the expression for this decrease, one starts from Eq. (3.60) for the amplitude for a wave scattered by a crystal. The positions of atoms may be given by their equilibrium position $\mathbf{r}_0$ and a time-dependent term $\mathbf{u}(t)$:

$$\mathbf{r}(t) = \mathbf{r}_0 + \mathbf{u}(t) \tag{3.73}$$

We assume that each atom oscillates around its equilibrium position completely independently. Then, the temperatures average yields

$$\begin{aligned}\left\langle \sum_j f_j \exp((\mathbf{r}_0 + \mathbf{u}(t))\mathbf{S}) \right\rangle &= \left\langle \sum_j f_j \exp 2\pi i \mathbf{r}_0\mathbf{S} \cdot \exp 2\pi i \mathbf{u}(t)\mathbf{S} \right\rangle \\ &= \left\langle \left(\sum_j f_j \exp 2\pi i \mathbf{r}_0\mathbf{S} \right) \right\rangle \cdot \langle \exp 2\pi i \mathbf{u}(t)\mathbf{S} \rangle \end{aligned} \tag{3.74}$$

where $\mathbf{S}$ is equal to the change of the wave vector during the reflection and the angle bracket $\langle \ldots \rangle$ denotes the thermal average. The first expression equivalent to Eq. (3.60) for the equilibrium expression $\mathbf{r}_0$ ensures the sharpness of all dif-

fracted reflections. The second term causes a decrease of its intensity. To show this, one develops it into a potential series:

$$\langle \exp 2\pi i \mathbf{u}(t)\mathbf{S}\rangle = 1 + i2\pi\langle \mathbf{u}(t)\mathbf{S}\rangle - \frac{1}{2}\left\langle 4\pi^2(\mathbf{u}(t)\mathbf{S})^2\right\rangle + \dots \tag{3.75}$$

However, $\langle \mathbf{u}(t)\mathbf{S}\rangle = 0$ because $\mathbf{u}(t)$ is a random movement, which is in no way correlated with the direction of $\mathbf{S}$. Furthermore, it holds that:

$$\langle (\mathbf{u}(t)\mathbf{S})^2\rangle = \frac{1}{3}\langle u(t)^2\rangle(\mathbf{S})^2 \tag{3.76}$$

The factor $\frac{1}{3}$ arises as, during the geometrical averaging in three dimensions, the component of the vector $\mathbf{u}$ in the direction of $\mathbf{S}$ plays a role. Thus, we obtain

$$\langle \exp 2\pi i \mathbf{u}(t)\mathbf{S}\rangle = 1 - \frac{1}{6}4\pi^2\langle u(t)^2\rangle(\mathbf{S})^2 + \dots \ . \tag{3.77}$$

It is quite useful to note that the function

$$\exp\left[-\frac{1}{6}4\pi^2\langle u(t)^2\rangle(\mathbf{S})^2\right] = 1 - \frac{1}{6}4\pi^2\langle u(t)^2\rangle(\mathbf{S})^2 + \dots \tag{3.78}$$

in its series expansion is identical in both first members to that of Eq. (3.77). This identity can be proved for all members for a harmonic oscillator. By substituting $(\mathbf{S})^2 = \dfrac{4\sin^2\theta}{\lambda^2}$ we get

$$\mathbf{F}(\mathbf{S}) = \mathbf{F}_0(\mathbf{S})\exp[-B(\sin^2\theta/\lambda^2)] \tag{3.79}$$

with $\mathbf{F}_0(\mathbf{S})$ the scattering amplitude of the rigid lattice and

$$B = \frac{8}{3}\pi^2\langle u(t)^2\rangle \tag{3.80}$$

The isotropic temperature factor as in this model the thermal motion has been assumed to be isotropic. In molecules, this is usually not the case and the thermal motion is described by a tensor ellipsoid with six independent parameters: three of these represent the dimensions of the principal axes, and three the orientation of these axes. The symmetric U tensor contributes to the factor, which is responsible for the temperature-dependence of the scattering, in the following way:

$$\begin{aligned}\exp[&- 2\pi^2(U_{11}h^2a^{*2} + U_{22}k^2b^{*2} + U_{33}l^2c^{*2} + 2U_{12}hka^*\,b^*\cos\gamma^* \\ &+ 2U_{13}hla^*\,c^*\cos\beta^* + 2U_{23}klb^*\,c^*\cos\alpha^*)]\end{aligned} \tag{3.81}$$

In protein crystallography, isotropic B values for each atom of the molecules are used normally. The thermal motion of the atoms is one main reason for the

fall-off in diffraction intensity, especially at higher diffraction angles. This limits the possible recordable number of diffraction spots and, as will be seen later, the resolution of the diffraction experiment. However, due to improved synchrotron radiation techniques and crystal quality there is a growing number of examples where the diffraction extends to real atomic resolution. In such cases the thermal parameters of the atoms are refined anisotropically.

3.2.8
Symmetry in Diffraction Patterns

An X-ray diffraction data set from a crystal represents its reciprocal lattice with the corresponding diffraction intensities at the reciprocal lattice points (*hkl*). As the reciprocal lattice is closely related to its direct partner, it reveals symmetries, lattice properties and other peculiarities (e.g., systematic extinctions) that are connected to the direct crystal symmetry, such as unit cell dimensions and space group. A detailed discussion of this problem is provided by Buerger (1961).

In the case of real atomic scattering factors f the diffraction intensities are centrosymmetric according to Friedel's law (Eq. 3.82):

$$I(hkl) = I(\overline{hkl}) \tag{3.82}$$

This is illustrated in Figure 3.12a and b. The square of a complex number is the product of this number by its complex conjugate. This is shown for $\mathbf{F}(hkl)$ in Figure 3.12a and for $\mathbf{F}(\overline{hkl})$ in Figure 3.12b. The resulting intensities are equal in both cases.

3.2.9
Electron Density Equation and Phase Problem

An inspection of the equation for the structure factor:

$$\mathbf{F}(\mathbf{S}) = \sum_{j=1}^{N} f_j \exp 2\pi i \mathbf{r}_j \mathbf{S} = \int_{\text{vol. of unit cell}} \rho(\mathbf{r}) \exp 2\pi i \mathbf{r}\mathbf{S} \, \mathrm{d}v \tag{3.83}$$

shows that it is the Fourier transform (FT) of the electron density $\rho(\mathbf{r})$.

The FT is of general relevance in broad areas of physics, for example, in optics, X-ray diffraction, FT-NMR techniques, and FT-IR methods. It has the following form as exponential FT:

$$\mathbf{F}(\mathbf{r}^*) = \int_s \rho(\mathbf{r}) \exp(2\pi i \mathbf{r} \cdot \mathbf{r}^*) d\mathbf{r} \tag{3.84}$$

$\mathbf{r}^*$ – vector in space of FT, $\mathbf{r}$ – vector in "direct space".

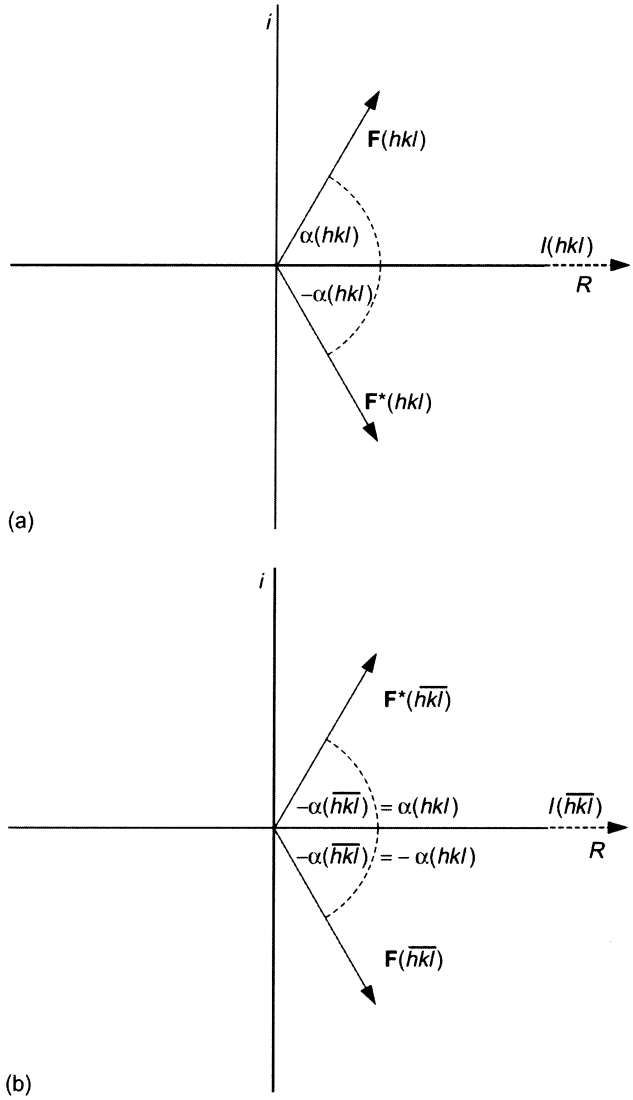


Fig. 3.12 Diagram illustrating the basis of Friedel's law.

It must be shown that

$$\rho(\mathbf{r}) = \int_{s^*} \mathbf{F}(\mathbf{r}^*) \exp(-2\pi i \mathbf{r}^* \cdot \mathbf{r}) d\mathbf{r}^* \,. \tag{3.85}$$

The right side of Eq. (3.85) becomes:

$$\int_{s^*} \left(\int_{s} \rho(\mathbf{r}') \exp(2\pi i \mathbf{r}^* \cdot \mathbf{r}') d\mathbf{r}' \right) \exp(-2\pi i \mathbf{r}^* \cdot \mathbf{r}) d\mathbf{r}^* \tag{3.86}$$

One can include the exponential functions and invert the integral signs:

$$\int_{s^*} \left(\int_{s} \rho(\mathbf{r}') \exp[2\pi i \mathbf{r}^* (\mathbf{r}' - \mathbf{r})] d\mathbf{r}' \right) d\mathbf{r}^* \tag{3.87}$$

$$\int_{s} \rho(\mathbf{r}') \left(\int_{s^*} \exp[2\pi i \mathbf{r}^* (\mathbf{r}' - \mathbf{r})] d\mathbf{r}^* \right) d\mathbf{r}' \tag{3.88}$$

The integral expression in parenthesis in Eq. (3.88) is known as a delta function $\delta(\mathbf{r}' - \mathbf{r})$. This has the following properties:

$$\delta = 0 \quad \text{for} \quad \mathbf{r}' \neq \mathbf{r}; \; \delta = \infty \quad \text{for} \quad \mathbf{r}' = \mathbf{r}; \; \int_{-\infty}^{\infty} \delta(\mathbf{r}' - \mathbf{r}) d\mathbf{r}' = 1 \; .$$

Hence we get:

$$\int_{s} \rho(\mathbf{r}') \delta(\mathbf{r}' - \mathbf{r}) d\mathbf{r}' = \rho(\mathbf{r}) \; . \tag{3.89}$$

One can also say:

$$\mathbf{F}(\mathbf{r}^*) = \mathrm{T}[\rho(\mathbf{r})] \tag{3.90}$$

and

$$\rho(\mathbf{r}) = \mathrm{T}^{-1}[\mathbf{F}(\mathbf{r}^*)] \tag{3.91}$$

with $\mathbf{F}$ the FT from ρ and ρ the FT from $\mathbf{F}$.

The electron density $\rho(\mathbf{r})$ is then the inverse FT of the structure factor $\mathbf{F}(\mathbf{S})$ according to Eq. (3.84):

$$\rho(\mathbf{r}) = \int_{\text{vol. of diffraction space}} \mathbf{F}(\mathbf{S}) \exp -2\pi i \mathbf{r}\mathbf{S} \mathrm{d}v_s \tag{3.92}$$

The integration is replaced by summation since $\mathbf{F}(\mathbf{S})$ is not continuous and is non-zero only at the reciprocal lattice points. Hence, we have:

$$\rho(xyz) = \frac{1}{V}\sum_{h=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty}\mathbf{F}(hkl)\cdot\exp[-2\pi i(hx+ky+lz) \tag{3.93}$$

By knowing the structure factors:

$$\mathbf{F}(hkl) = |\mathbf{F}(hkl)|\exp ia(hkl) \tag{3.94}$$

one can calculate the electron density distribution in the unit cell and thus determine the atomic positions of the scattering molecule(s). Unfortunately, the measured quantities are only the absolute values $|\mathbf{F}(hkl)|$ of the structure factor. Information on the phase angles $a(hkl)$ is lost during the diffraction experiment. The determination of these phases is the basic problem in any crystal structure determination, and methods for solving the phase problem are discussed later.

3.2.10
The Patterson Function

The measured X-ray intensities are proportional to the square of the absolute value of the structure factor according to Eq. (3.64). Would it be possible to use the intensities directly to calculate from these a function which contains structural information? The answer is "yes". By calculating a convolution of the electron density with itself, Patterson (1934) showed that this is just the FT of the intensities (Eq. 3.95):

$$P(\mathbf{u}) = \int_{\text{vol. of unit cell}} \rho(\mathbf{x})\rho(\mathbf{x}+\mathbf{u})d\mathbf{x} \tag{3.95}$$

To prove this, we substitute $\rho(\mathbf{x})\frac{1}{V} = \sum_{\mathbf{h}=-\infty}^{\infty}\mathbf{F}_{\mathbf{h}}\exp[-2\pi i\mathbf{hx}]$ and $\rho(\mathbf{x}+\mathbf{u}) = \frac{1}{V}\sum_{\mathbf{h}'=-\infty}^{\infty}\mathbf{F}_{\mathbf{h}'}\exp[-2\pi i\mathbf{h}'\cdot(\mathbf{x}+\mathbf{u})]$ with the position vector $\mathbf{x}$ with components (x, y, z) and the reciprocal vector $\mathbf{h}$ with the integral triples of numbers (h, k, l) and get for the summation over $\mathbf{h}$:

$$P(\mathbf{u}) = \frac{1}{V^2}\sum_{\mathbf{h}}\sum_{\mathbf{h}'}\mathbf{F}_{\mathbf{h}}\cdot\mathbf{F}_{\mathbf{h}'}\exp[-2\pi i\mathbf{h}'\cdot\mathbf{u}]\int_{\text{vol. of unit cell}}\exp[-2\pi i(\mathbf{h}+\mathbf{h}')\cdot\mathbf{x}]d\mathbf{x}\,. \tag{3.96}$$

Considering the integral

$$\int_{\text{vol. of unit cell}} \exp[-2\pi i(\mathbf{h}+\mathbf{h}')\cdot\mathbf{x}]d\mathbf{x}$$
$$= \int_0^a \exp\left[-2\pi i(h+h')\frac{X}{a}\right]dX \int_0^b \exp\left[-2\pi i(k+k')\frac{Y}{b}\right]dY \int_0^c \exp\left[-2\pi i(l+l')\frac{Z}{c}\right]dZ \tag{3.97}$$

where $\mathbf{h} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$, $\mathbf{h}' = h'\mathbf{a}^* + k'\mathbf{b}^* + l'\mathbf{c}^*$ and $\mathbf{x} = \frac{X}{a}\mathbf{a} + \frac{Y}{b}\mathbf{b} + \frac{Z}{c}\mathbf{c}$ were explicitly substituted and the relationships [Eq. (3.67)] between the direct and reciprocal lattices were applied. The integration of the individual integrals of the same type, exemplified with the left-handed integral, reveals:

$$\int_0^a \exp\left[-2\pi i(h+h')\frac{X}{a}\right]dX = \left[\frac{\exp\left[-2\pi i(h+h')\frac{X}{a}\right]}{-2\pi i(h+h')}\right]_0^a . \tag{3.98}$$

The exponential in the numerator becomes 1 because h and h' are integers and $X=a$ or 0. Therefore, the integral is, in general, zero. The same holds for k and k' and for l and l'. On the other hand, when $h = -h'$, $k = -k'$ and $l = -l'$, Eq. (3.97) degenerates directly into

$$\int_0^a \exp(0)dX \int_0^b \exp(0)dY \int_0^c \exp(0)dZ = \int_0^a\int_0^b\int_0^c dXdYdZ = V \tag{3.99}$$

From this it follows that [Eq. (51)]:

$$P(\mathbf{u}) = \frac{1}{V}\sum_{\mathbf{h}} F_{\mathbf{h}}^2 \exp -2\pi i\mathbf{hu} \tag{3.100}$$

Thereby it was taken into account that $\mathbf{F}_{\mathbf{h}} \cdot \mathbf{F}_{-\mathbf{h}} = F_{\mathbf{h}}^2$, which furthermore reduces the double sum in Eq. (3.96) into a single sum over $\mathbf{h}$.

The self-convolution of a one-dimensional electron density distribution of three atoms according to Eq. (3.95) is explained in Figure 3.13 a–c, which shows two repeats of the cell of length a. The centers of the atoms are at the maxima of $\rho(x)$ and between the atoms the electron density has negligible values. The possible interatomic vectors are indicated. The convolution means that $\rho(x)$ is shifted by u [e.g., $u = 0.7$ in Fig. 3.13 b and $u = 3.0$ in Fig. 3.13 c] and the product of $\rho(x)$ and $\rho(x+u)$ is integrated for x. If u has values that the maxima in both function do not coincide with (as in Fig. 3.13 b), the products of the relevant electron density values will have small values. However, in the case that u equals an interatomic

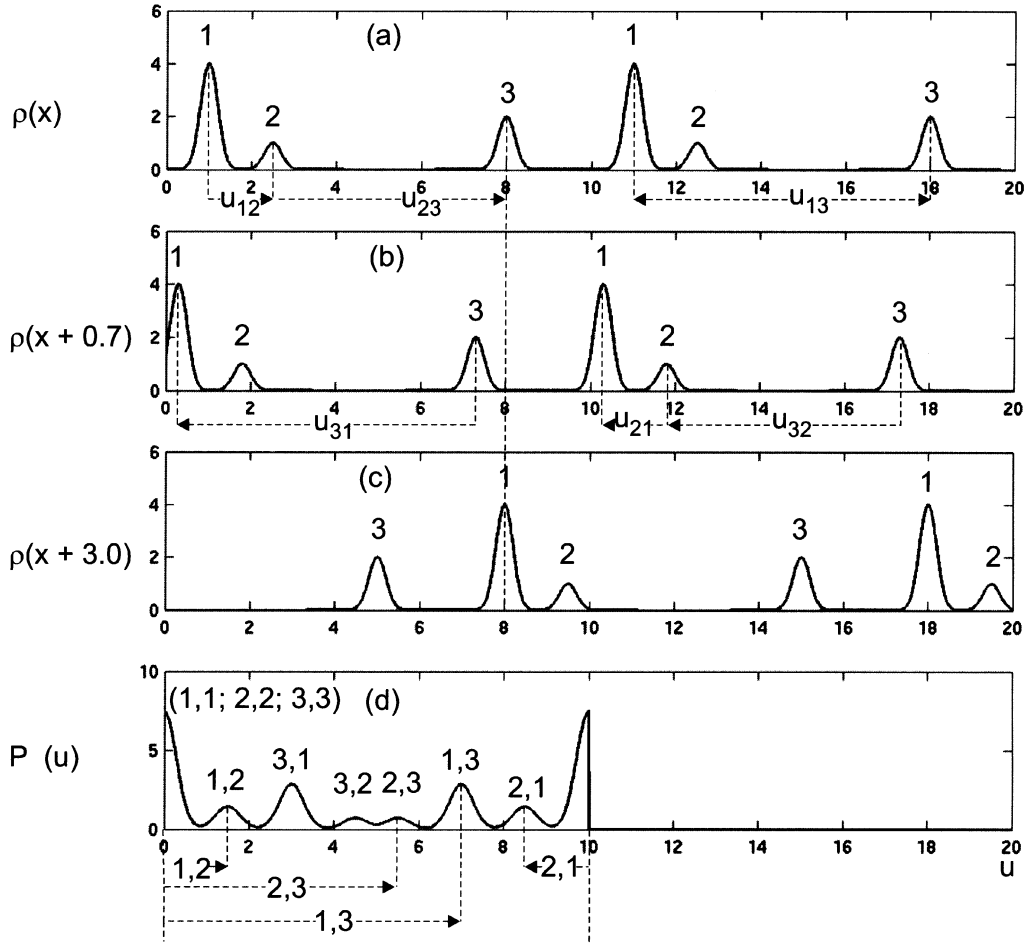


Fig. 3.13 Convolution of a one-dimensional electron density distribution with itself. (a) Three-atomic structure $\rho(x)$; (b) $\rho(x)$ shifted by $u = 0.7$; (c) $\rho(x)$ shifted by $u = 3.0$; (d) $P(u)$. (Figure was produced with MATCAB; Math Works, Inc., 2005.)

distance, the two relevant maxima coincide (as in Fig. 3.13 c for peaks 3 and 1), and the products of the electron densities around this position will reveal large values, giving a peak in the function $P(u)$. Figure 3.13 d shows the respective complete function with three self-vectors at the origin and six possible interatomic distances.

In three dimensions, the function $P(\mathbf{u})$ will have maxima if the positions $\mathbf{x}$ and $\mathbf{x} + \mathbf{u}$ correspond to atoms. In general, we obtain a function that contains the interatomic vectors as maxima. We expect N^2 peaks for N atoms. The maxima are proportional to $Z_i Z_j$.

The Patterson maxima of a three-atom structure in two dimensions are depicted in Figure 3.14. The distribution of the $N^2 = 9$ maxima (containing three

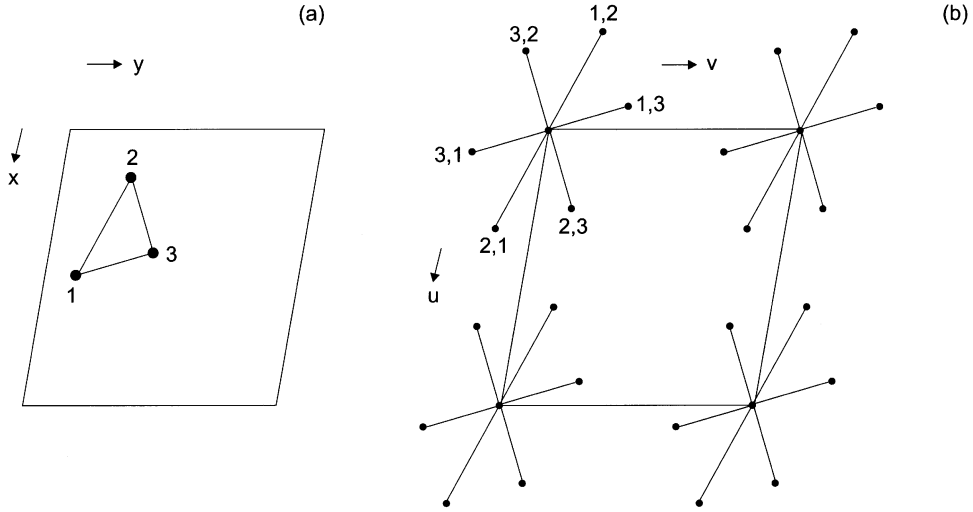


Fig. 3.14 (a) Two-dimensional three atom structure; (b) appropriate Patterson function.

self-vectors at the origin) is centro-symmetric, which means that for each vector u, v, w a vector $-u, -v, -w$ exists. The centro-symmetry is a general property of the Patterson function.

The Patterson function is a very useful tool to locate atoms when the number of atoms in the asymmetric unit of the unit cell is not too high (e.g., <20), or it contains a subset of heavy atoms among not too many (e.g., <100) light atoms such as C, N, O, or S. Here, the heavy atom–heavy atom vectors are clearly prominent. If a protein with 1000 (or a multiple of that) light atoms holds one or several heavy atoms per molecule, the signal resulting from the heavy atoms can no longer be resolved. However, when using the method of isomorphous replacement (discussed later), a Patterson function of the heavy atom structure can be calculated, from which it is possible to locate the heavy atoms.

3.2.11
Lorentz Factor and Integrated Intensity Diffracted by a Crystal

We have seen that the interference function I_F contributes to the scattering power in the immediate vicinity of the reciprocal lattice points. In order to obtain the whole scattering power originating from a particular reciprocal lattice point, we have to rotate it through the Ewald sphere and carry out an integration over $I(\mathbf{S})$ [Eq. (3.58.3)]. This is identical to a rotation of the Ewald sphere in the opposite direction, as shown in Figure 3.15. We assume that the rotation axis passes through O, the origin of the reciprocal lattice, and is normal to the incident beam, as in the rotation data collection technique. In position 1, the reflection sphere cuts the reciprocal lattice point P (strongly exaggerated in Fig. 3.15). The space angle element may be chosen sufficiently small that the

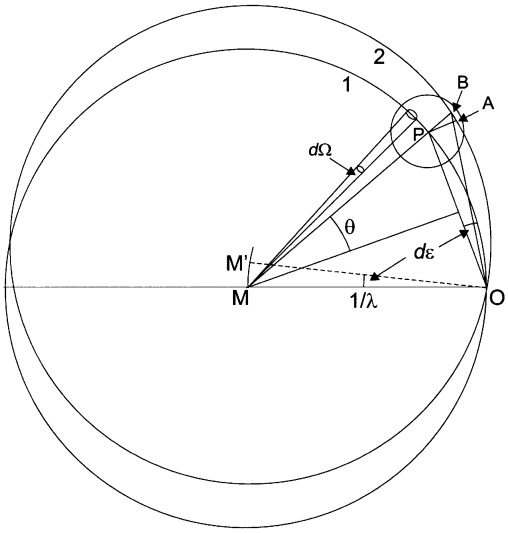


Fig. 3.15 Explanatory diagram for the derivation of the whole scattering power originating from a particular reciprocal lattice point.

flux $I(\mathbf{S})$ can be assumed as constant. If we run over from flux to power, the reflected beam power observed at the model point in distance r_0 becomes $I(\mathbf{S})r_0^2 d\Omega$. Thus, the whole power of the reflection sphere in position 1 observed at the model point will be

$$C_1(\mathbf{S}) = \int I(\mathbf{S}) r_0^2 d\Omega \,. \tag{3.101}$$

We determine $d\Omega$ for the first dimension as $d\Omega = \dfrac{\Delta \mathbf{r}^*}{1/\lambda} = \lambda dha^*$ and for two dimensions $d\Omega = \lambda^2 dhdka^*b^*$. Thus, we reveal

$$C_1(\mathbf{S}) = \lambda^2 r_0^2 \int\int I(\mathbf{S}) dhdka^*b^* \tag{3.102}$$

As the spherical area around P is very small, the tangent plane at the reflection sphere at point P will be well approximated by the $\mathbf{a}^*\mathbf{b}^*$ plane. Thus, we obtain the whole reflected power in the model point if the reflection sphere sweeps over the reciprocal point around P in incremental values of the rotation angle ε.

$$C(\mathbf{S}) = \lambda^2 r_0^2 \int\int\int I(\mathbf{S}) dhdka^*b^*d\varepsilon \tag{3.103}$$

We replace $d\varepsilon$ by the increment of l, dl. For this purpose, we introduce the third reciprocal lattice vector $\mathbf{c}^*$, which is normal to $\mathbf{a}^*, \mathbf{b}^*$ and parallel to line PB in Figure 3.15. Here, PA $= |S| d\varepsilon$ with $|S| = \dfrac{2\sin\theta}{\lambda}$ and PB $=$ PA $\cos\theta = \dfrac{\sin 2\theta}{\lambda} d\varepsilon$. We set c^*dl for PB and get with $d\varepsilon = \dfrac{\lambda}{\sin 2\theta} dlc^*$

$$C(\mathbf{S}) = \frac{\lambda^3 r_0^2}{\sin 2\theta} \int\int\int (\mathbf{S}) dh dk dl a^* b^* c^* \tag{3.104}$$

As the extension of the reciprocal lattice points is very small, it can be assumed that one reflection is in reflection position in a time only, and that the integration can range from $-\infty$ to $+\infty$. Equation (3.104) contains the interference function I_F with the product of the three integrals of the type

$$\begin{aligned} \int_{-\infty}^{+\infty} \frac{\sin^2 N_1(\pi\Delta h)}{\sin^2 \pi\Delta h} a^* dh &= \frac{a^*}{\pi} \int_{-\infty}^{+\infty} \frac{\sin^2 N_1(\pi\Delta h)}{\sin^2 \pi\Delta h} \pi dh \\ &\cong \frac{a^*}{\pi} \int_{-\infty}^{+\infty} \frac{\sin^2 N_1 u}{u^2} du = \frac{a^* N_1}{\pi} \int_{-\infty}^{+\infty} \frac{\sin^2 v}{v^2} dv = a^* N_1 \end{aligned} \tag{3.105}$$

Thus, the product of the three integrals reveals:

$$a^* b^* c^* N_1 N_2 N_3 = V^* N_1 N_2 N_3 \tag{3.106}$$

We obtain the number of unit cells $N_1 N_2 N_3$ of a crystal by dividing its volume V_x by the volume V of its unit cell. Taking into account that $V^* = \frac{1}{V}$, we obtain $\frac{V_x}{V^2}$ for Eq. (3.106). The integrated scattering power of a small crystal over the whole range around the reciprocal lattice point in consideration of Eq. (3.58.3) is then:

$$C_{\mathbf{h}} = \frac{e^4}{c^4 \mu^2} \left(\frac{1 + \cos^2 2\theta}{2} \right) \frac{\lambda^3}{\sin 2\theta} I_0 |F_{\mathbf{h}}|^2 \frac{V_x}{V^2} \tag{3.107}$$

The described integration has been carried out by H.A. Lorentz, and follows the description provided by Wölfel (1987). The factor $\frac{1}{\sin 2\theta} = L$ is characteristic for the treated rotation geometry and is denoted as the "Lorentz factor". The power C is related to the total energy E under the reflection curve by $C = E\omega$ with ω, the constant angular velocity of the rotation of the reciprocal lattice point through the Ewald sphere. E is now a quantity to be measured. Hence, we obtain:

$$E(\mathbf{h}) = \frac{I_0}{\omega} \lambda^3 \frac{e^4}{c^4 \mu^2} p \frac{LAV_x}{V^2} |\mathbf{F}(\mathbf{h})|^2 \tag{3.108}$$

Here, p=polarization factor, L=Lorentz factor (a geometrical factor taking into account the relative time that each reflection spends in the reflection position) and A=absorption factor have been introduced. Equation (3.108) is also known as Darwin's equation.

The polarization factor p for unpolarized radiation is $\frac{(1+\cos^2\theta)}{2}$, as in Eq. (3.107). It adopts different values for radiation emerging from a monochromator or for synchrotron radiation, which is strongly polarized.

The absorption of an X-ray beam with intensity I_0 traveling through matter with a path length of t leads to a reduction in intensity according to

$$I = I_0 \exp[-\mu t] \tag{3.109}$$

where μ is the total linear absorption coefficient. Absorption is mainly generated by two effects:

- Photoelectric absorption, where the absorption may become substantially stronger if the X-ray photon can strike out an electron of the atom. This is the case if the energy of the X-ray photon is around the absorption edge of the atom. The effect is then denoted as anomalous scattering and is exploited when determining the phases of the structure factors (see Section 5.2).
- Scattering: the X-ray photon is scattered out of its primary direction either with energy loss (Compton scattering) or without loss (Rayleigh scattering).

Until now, we have assumed that I_0 is constant within the crystal (kinematic theory of X-ray diffraction). Ordinary absorption effects have been considered in Eq. (3.108). Experimental data have shown that Eq. (3.108) represents the total energy of a reflected beam very well, but for what reason? Real crystals are not perfect; rather, they should be regarded as consisting of small blocks of perfect crystals (of size ca. 0.1 μm) which have an average tilt angle among each other of 0.1–0.5° for protein crystals, and diffract independently of each other. Such a real crystal is denoted a *mosaic* crystal.

Owing to this mosaicity (0.1–0.5°), each reflection has a corresponding reflection width which is much larger than that originating from the interference function I_F. The integrated intensity equation is valid because the mosaic blocks are so small that no multiple scattering occurs within an individual mosaic block, and the attenuation of the primary intensity I_0 is negligible due to regular Bragg reflection. The model of the mosaic crystal also explains why no Umweg excitation occurs, although several reciprocal lattice points may be in reflection position at the same time due to the large unit cells in protein crystals.

The integrated intensity depends on λ to the third power. Increasing the wavelength causes appreciably stronger diffraction intensities, but this is accompanied by greater absorption. Cu Kα radiation with a wavelength of 1.5418 Å is an optimal choice for protein crystallography when using X-ray generator sources. Also important is the dependence of the integrated intensity on the unit cell volume V by its negative second power. Doubling of the unit cell volume with twice as many molecules, taking into account the increase in $|\mathbf{F}(\mathbf{h})|^2$ by having now $2n$ molecules per unit cell, reduces the average intensity for the reflected beams by a factor of two.

In Eq. (3.108) $(\lambda^3/\omega V^2)\cdot(e^4/c^4)\cdot V_x \cdot I_0$ is a constant for a given experiment. The corrected intensity on a relative scale $I(\mathbf{h})$ is obtained from:

$$I(\mathbf{h}) = \frac{E(\mathbf{h})}{p \cdot L \cdot A} \tag{3.110}$$

3.2.12
Intensities on an Absolute Scale

The corrected intensity on a relative scale $I(\mathbf{h})$ can be converted to an intensity given by:

$$I(abs, \mathbf{h}) = \mathbf{F}(\mathbf{h})\mathbf{F}(\mathbf{h})^* = |\mathbf{F}(\mathbf{h})|^2 \tag{3.111}$$

on an absolute scale by applying a so-called Wilson plot. The basis for this plot is an equation which connects the average intensity on an absolute scale with the average intensity on a relative scale by a scale factor C, and considers the isotropic thermal motion of the scattering atoms by the temperature factor given in Eq. (3.80). This is written in the form of:

$$\ln \frac{\overline{I(\mathbf{h})}}{\sum_j (f_j)^2} = \ln C - 2B\frac{\sin^2\theta}{\lambda^2} \tag{3.112}$$

This is the equation of a straight line. B, the overall temperature factor, and C, the scale factor, can be obtained by plotting $\ln \overline{I(\mathbf{h})}/\sum_j (f_j)^2$ against $(\sin^2\theta)/\lambda^2$.

3.2.13
Resolution of the Structure Determination

The concept of resolution in X-ray diffraction has the same meaning as the concept in image formation in the optical microscope.

After the Abbe theory, we obtain:

$$d_m = \frac{\lambda}{2NA} \tag{3.113}$$

where NA is the numerical aperture of the objective lens. In protein crystallography, the nominal resolution of an electron density map is expressed in d_m, the minimum interplanar spacing for which Fs are included in the Fourier series. The maximum attainable resolution at a given wavelength is $\lambda/2$. For copper Kα radiation it is 0.7709 Å, and this would suffice to determine protein structures at atomic resolution (the distance of a carbon–carbon single bond is about 1.5 Å). However, the thermal vibrations of the atoms in a protein crystal are usually so high that the diffraction data cannot be observed to the full theoretical resolution limit. The polypeptide chain fold can be determined at a resolution of better than 3.5 Å. A medium-resolution structure is in the resolution

range of 3.0–2.2 Å, which makes the amino acid side chains clearly visible. A high-resolution structure has a nominal resolution better than 2.2 Å, and can be as good as 1.2 Å. In such structures the main-chain carbonyl oxygens become visible as prominent bumps, and at a resolution better than 2.0 Å aromatic side chains acquire a hole in the middle of their ring systems. For some very well diffracting crystals from small proteins, diffraction data extending to resolutions below 1.2 Å could be collected with synchrotron radiation (Wilson, 1998). Such structures reveal real atomic resolution where each atom is visible as an isolated maximum in the electron density map.

References

Blundell, T. L., Johnson, L. N., *Protein Crystallography*, Academic Press, New York, **1976**.

Buerger, M. J., *Crystal-structure Analysis*, John Wiley & Sons, New York, **1961**.

Patterson, A. L., *Phys. Rev.* **1934**, *46*, 372–376.

Prince, E. (Ed.), *International Tables for Crystallography, Volume C: Mathematical, physical and chemical Tables*, Springer, New York, **2004**.

Von Laue, M., *Röntgenstrahlinterferenzen*, Akademische Verlagsgesellschaft Geest & Portig KG, Leipzig, **1948**.

Wilson, K. S., *Nat. Struct. Biol. Synchrotron Suppl.* **1998**, pp. 627–630.

Wölfel, E. R., *Theorie und Praxis der Röntgenstrukturanalyse. Eine Einführung für Naturwissenschaftler*, Vieweg Friedrich & Sohn Verlagsgesellschaft, Stuttgart, **1987**.

4
Diffraction Data Evaluation

4.1
Introductory Remarks

The analysis and reduction of diffraction data from a single crystal consists of seven main steps:

1. Visualization and preliminary analysis of the raw, unprocessed data.
2. Indexing of the diffraction patterns.
3. Refinement of the crystal and detector parameters.
4. Integration of the diffraction spots.
5. Finding the relative scale factors between measurements.
6. Precise refinement of crystal parameters using the whole data set.
7. Merging and statistical analysis of the measurements related by space-group symmetry.

When using electronic area detectors with short read-out times such as charge-coupled device (CCD) or multiwire proportional chamber (MWPC) detectors it is possible to collect diffraction images with small rotational increments (0.05–0.2°). In this case, the reflection profile over the crystal rotation angle can be registered, giving a three-dimensional picture of the spot. The evaluation of such diffraction data can be made with computer programs MADNES (Messerschmidt and Pflugrath, 1987), XDS (Kabsch, 1988a, b, 1993), the San Diego programs (Howard et al., 1985) and related programs XENGEN (Howard et al., 1987) and X-GEN. IP systems with their longer read-out times are operated in a film-like mode with rotational increments of 0.5 to 2.0°. Here, mainly the program systems MOSFLM (Leslie, 1999) and HKL-2000 (Otwinowski and Minor, 1997) are applied. However, these programs are now able also to handle rotational increments of smaller values (e.g., 0.1°), as are often applied with CCD-detectors at synchrotron PX beamlines. Currently, MOSFLM, HKL-2000 and XDS are the most popular data evaluation programs.

4.2 Geometric Principles in the Rotation Technique with Normal Flat Detector

Almost all X-ray diffraction data collection set-ups in protein crystallography apply the rotation technique, with a flat detector normal to the incident primary beam. This can be one of the area detectors treated in Section 2.2. Figure 4.1 shows the geometric relationships for the rotation technique with a flat detector normal to the primary X-ray beam. The definition of the coordinate systems is those used in the data evaluation program MOSFLM (Leslie, 1999). The origin of the Ewald sphere and of the ortho-normal laboratory coordinate system X, Y, Z is in C. The flat detector is in distance D with the ortho-normal coordinate system X_d, Y_d centered at O′. The rotation axis of the crystal is parallel to the Z-axis, and the sense of rotation is indicated. The origin of the detector pixel coordinate system X_S, Y_S is at O_S. A reciprocal lattice point P is in reflection position in Figure 4.1, and is recorded at P′ on the detector. The reciprocal lattice point P has the coordinates x, y, z (**x** as vector) in the ortho-normal coordinate system x, y, z of the reciprocal lattice with its origin in O. We obtain its coordinates from Miller's indices h, k, l (**h** as vector) by Eq. (4.1):

$$\mathbf{x} = \Phi \times \mathbf{PHIZ} \times \mathbf{PHIY} \times \mathbf{PHIX} \times \mathbf{A} \times \mathbf{h} \tag{4.1}$$

with $\mathbf{A} = \begin{pmatrix} a_x^* & b_x^* & c_x^* \\ a_y^* & b_y^* & c_y^* \\ a_z^* & b_z^* & c_z^* \end{pmatrix}$ the matrix of the components of the reciprocal cell vectors with respect to the coordinate system x, y, z.

$\mathbf{PHIX} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi_x & -\sin\varphi_x \\ 0 & \sin\varphi_x & \cos\varphi_x \end{pmatrix}$, **PHIY** and **PHIZ** are the corresponding rotations about the x-, y-, and z-axes, respectively. φ_x, φ_y, φ_z are designated as missetting angles because they correct for rotational movements of the crystal during the data collection. The action of the first four matrices brings the crystal in the starting position of the crystal rotation axis with $\varphi = 0$. The rotation of the crystal about this spindle axis is taken into account by matrix Φ.

The coordinates x, y, z of P on the reflection sphere can be derived from the detector coordinates according to Figure 4.1. Thereby, intercept theorems are applied. The following relationships hold:

$$\mathrm{O'P'} = \sqrt{X^2 + Y^2},\ \mathrm{CP'} = \sqrt{X^2 + Y^2 + D^2}\ . \tag{4.2}$$

$$\frac{\mathrm{NP}}{\mathrm{NP'}} = \frac{\mathrm{CP}}{\mathrm{CP'}} = \frac{y}{X} = \frac{1/\lambda}{\sqrt{X^2 + Y^2 + D^2}},\ y = \frac{X}{\lambda\sqrt{X^2 + Y^2 + D^2}} \tag{4.3}$$

$$\frac{\mathrm{MN}}{\mathrm{O'N'}} = \frac{\mathrm{CN}}{\mathrm{CN'}} = \frac{\mathrm{CP}}{\mathrm{CP'}} = \frac{z}{Y} = \frac{1/\lambda}{\sqrt{X^2 + Y^2 + D^2}},\ z = \frac{Y}{\lambda\sqrt{X^2 + Y^2 + D^2}} \tag{4.4}$$

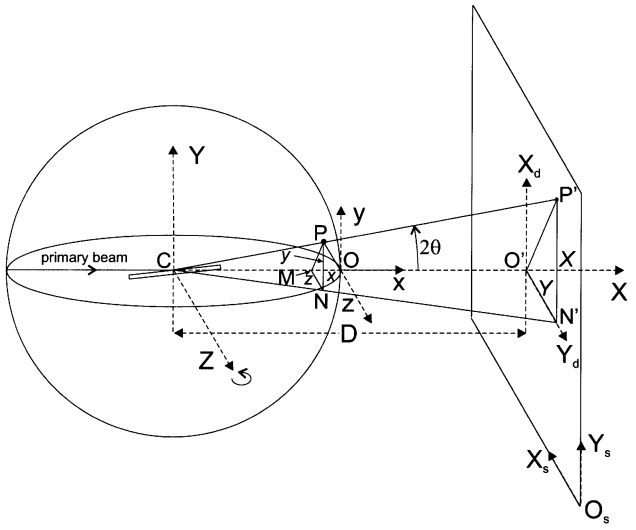


Fig. 4.1 Geometric relationship for the rotation technique with a flat detector normal to the primary X-ray beam.

$$\frac{\mathrm{CM}}{\mathrm{CO}'} = \frac{\mathrm{CP}}{\mathrm{CP}'} = \frac{1/\lambda - x}{\mathrm{D}} = \frac{1/\lambda}{\sqrt{X^2 + Y^2 + \mathrm{D}^2}},\ x = \frac{1}{\lambda}\left(\frac{\mathrm{D}}{\sqrt{X^2 + Y^2 + \mathrm{D}^2}}\right) \tag{4.5}$$

The detector coordinates X, Y are obtained from the following relationships:

$$\frac{X}{y} = \frac{\mathrm{D}}{1/\lambda - x},\ X = \frac{\mathrm{D}y}{1/\lambda - x} \tag{4.6}$$

$$\frac{Y}{z} = \frac{\mathrm{D}}{1/\lambda - x},\ Y = \frac{\mathrm{D}z}{1/\lambda - x}\ . \tag{4.7}$$

The MOSFLM system defines the following camera constants, which must be supplied by the user, and calculates corrections for these. XCEN, YCEN are the coordinates of the beam center in the scanner coordinate system. Any deviations in the refined position of the center of the diffraction pattern from these coordinates are denoted by the camera constants CCX and CCY. Any deviation of the angle OMEGA from its expected value of 90° is referred to by the camera constant, CCOM. The camera constants allow for errors in the user-defined position of the direct beam (CCX, CCY) and in the alignment of the detector pixel coordinate system relative to the camera (and detector) axes (CCOM).

If an approximate matrix **A** is known, the Miller indices of an observed peak at (X, Y) can be approximately determined using Eqs. (4.3)–(4.6) and (4.1)

$$\mathbf{h} = \mathbf{A}^{-1} \times \boldsymbol{\Phi}^{-1} \times \mathbf{x} \tag{4.8}$$

Hereby, the starting matrices of the inverse matrices of **PHIX**, **PHIY**, and **PHIZ** are the unity matrices and omitted in Eq. (4.8). The Miller indices have errors which depend upon the width of the oscillation range, the error in the detector parameters, and errors in determining the coordinates of the centers of the recorded reflections.

4.3 Autoindexing of Oscillation Images

Autoindexing routines have been successfully used in single crystal diffractometry with point detectors for initiating data collection (Sparks, 1976, 1982). These methods are based on accurately determined reciprocal lattice vectors for a few selected reflections. A challenging task was the autoindexing of oscillation images of macromolecules recorded on 2D-detectors using randomly oriented crystals. Indexing of such images without any prior knowledge is important for several reasons:

- most crystals are frozen in a cryo-loop for data collection and usually have a random orientation;
- the life time of a crystal in an X-ray beam is limited, and prealignment experiments would shorten its life time or would unnecessarily prolong the user time at a synchrotron PX beamline;
- the completeness of a diffraction data set is much higher if the crystal is rotated through the reflection sphere in a random orientation. This problem was solved recently, with the developed methods applying a Fourier analysis of one-dimensional distributions of observed reciprocal lattice points projected onto a chosen direction. This is used in the program DENZO, a part of the HKL-2000 package (Otwinowski and Minor, 1997) and in program DPS (Steller et al., 1997), which has been integrated into MOSFLM (Leslie, 1999).

Here, the method of program DPS is explained in some detail in order to understand its principles, using Figure 4.2 as an illustration. The reciprocal lattice points, which lie between the two circular arcs of the reflection sphere rotated around the oscillation angle $\Delta\varphi$, give rise to reflections that are recorded on the detector. Their coordinates **x** in the coordinate system x, y, z centered in O can be calculated from the detector coordinates *X*, *Y* by Eqs. (4.3) to (4.5). Now, each reciprocal lattice point is projected onto a chosen unit vector **t** with polar coordinates ψ and φ, as indicated for $\mathbf{t}_1$ in Figure 4.2. The projection p is then:

$$p = \mathbf{x} \cdot \mathbf{t} \tag{4.9}$$

We must sample all such projections of the reciprocal lattice points onto the given direction **t** for the application of a discrete fast-Fourier transform (FFT) algorithm and obtain the frequencies $f(p)$. An experimental frequency distribution with easily recognizable periodic distributions is depicted in Figure 4.3.

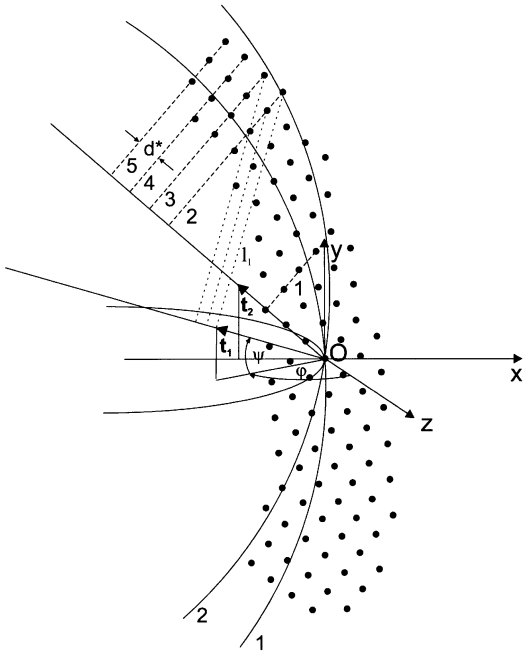


Fig. 4.2 Projection of reciprocal lattice vectors which are in reflection position onto two different direction vectors $\mathbf{t}_1$ and $\mathbf{t}_2$.

Then, the frequency $f(p)$ in the range $p \leq \mathbf{x} \cdot \mathbf{t} \leq p + \Delta p$ can be given by $f(p)\Delta p = f(j)$, where j is the closest integer to $(p - p_{\min})/\Delta p$ and $\Delta p = n a_{\max}$ with $p_{\min}$, the minimal value of p for the given direction, $a_{\max}$, maximal real cell dimension and n, number of grid points between successive reciprocal lattice planes. Thus, the discrete FT of this frequency distribution will be given by the summation

$$F(k) = \sum_{j=0}^{m} f(j) \exp(2\pi i k j) \tag{4.10}$$

with m the number of grid points along direction $\mathbf{t}$ and is calculated using a FFT algorithm between 0 and $m/2$ (Fig. 4.4). In Figure 4.2, two direction vectors $\mathbf{t}$ are shown; vector $\mathbf{t}_1$ has such a position that the projections (e.g., lines l_1) of the reciprocal lattice points reveal frequencies $f(p)$ of 1 or 2 only. The situation is different when $\mathbf{t}$ is perpendicular to a reciprocal lattice plane, as $\mathbf{t}_2$ in Figure 4.2. The frequency is 1 for line 1, but 3 for lines 2 to 5. Thus, the Fourier coefficients that best represent the periodicity will be large. The largest coefficient will occur at $k = 0$ and represent the number of vectors used in calculating the frequency distribution. The next set of large coefficients will correspond to the

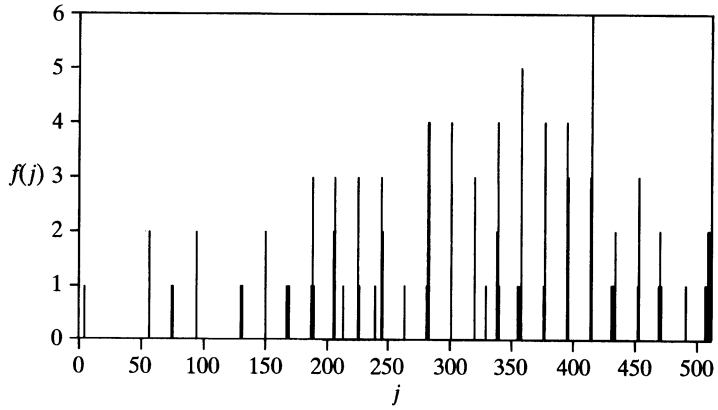


Fig. 4.3 Frequency distribution of the reciprocal lattice vectors for a suitable chosen direction of a diffraction pattern from a fibritin crystal. (Reproduced with permission from Steller et al., 1997; International Union of Crystallography.)

periodicity that represents every reciprocal lattice plane. Subsequent maxima will be due to periodicities spanning every second, third, etc. frequency maximum, and will thus be progressively smaller (Fig. 4.4). The largest $F(k)$ (when $k = l$), other than $F(0)$, will, therefore, correspond to an interval of d^* between reciprocal lattice planes in the direction of **t** where $d^* = l/(na_{\max})$.

The Fourier analysis is carried out for each direction **t** in a range from $0 \leq \psi \leq \pi/2$, $0 \leq \varphi \leq 2\pi$ and the relevant $F(k)$ coefficients related to the largest local maximum at $k = l$ of each direction **t** are determined. A set of the ψ and φ

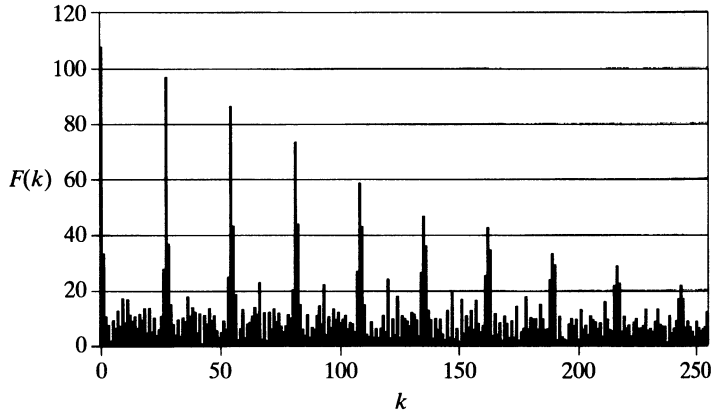


Fig. 4.4 Fourier analysis of the distribution shown in Figure 4.3. The first maximum, other than $F(0)$, is at $k = 27$, corresponding to $(1/d^*) = 41.9$ Å and a value of $F(27) = 97.0$. (Reproduced with permission from Steller et al., 1997; International Union of Crystallography.)

values associated with the largest maxima are subjected to a refinement procedure. The $F(l)$ values of the refined positions are then sorted by length, and a linearly independent set of three basis vectors of a primitive real space unit cell is chosen. As shown in Figure 4.2, the determined periodicities correspond to the plane distances of the reciprocal lattice d^*. Its inverse is a vector in direct space normal to the stack of planes of the reciprocal lattice. They must be converted to the basis vectors of the reciprocal cell to obtain the components of the three reciprocal cell axes along the three camera axes, which are the nine coefficients of the crystal orientation matrix **A**. Various nonlinear combinations of the refined vectors with largest $F(l)$ values are selected and used for a test indexing of the diffraction image. In most cases the best combination conforms to taking the three largest $F(l)$ values.

Finally, the reduced cell is determined for the best cell obtained. This cell is then analyzed in terms of 44 lattice characters (Burzlaff et al., 1992; Kabsch, 1993) in order to evaluate the most probable Bravais lattice and crystal system.

4.4
Beam Divergence, Mosaicity, and Partiality

Crystals of biological macromolecules have in general large cell constants (about 100 Å and larger), which means that the reciprocal lattice planes are densely populated with reciprocal lattice points. The Ewald construction tells us what reflection pattern is generated for a stack of reciprocal lattice planes in a given orientation with respect to the incident X-ray beam. If the crystal is stationary during the exposure a so-called "still" image is obtained. Such an image is shown in Figure 4.5, where the stack of reciprocal lattice planes is nearly normal to the primary beam with a flat detector perpendicular to the incident beam. The planes cut the Ewald sphere in circles. The reflected beams lie on cones, with the apex in the center of the Ewald sphere. The intersection of these cones with the detector reveals ellipses, as depicted in Figure 4.5. It is now observed on experimental images that the width of reflections and of the ellipses vary considerably. This is due to three factors. First, it is assumed that the primary X-ray beam is totally parallel. This is not the case with real X-ray sources, which owe a beam divergence δ. Second, real crystals are not perfect and exhibit a mosaicity η, which is below 0.05° for good crystals at room temperature but may increase to more than 1° due to crystal freezing or bad crystal quality. In reciprocal space, this corresponds to enlargement of the reciprocal lattice points with respect to their size originating from the interference function I_F (as discussed in Chapter 3). Third, X-radiation is only monochromatic within a defined wavelength bandpass $\delta\lambda/\lambda$, in the range of 0.0002–0.001 at synchrotron beamlines, but is considerably larger for laboratory sources. The bandpass, in effect, thickens the surface of the reflection sphere.

These effects are illustrated schematically in Figure 4.6. The combined result is that the diffraction of a particular reflection is spread over a range of crystal rotation.

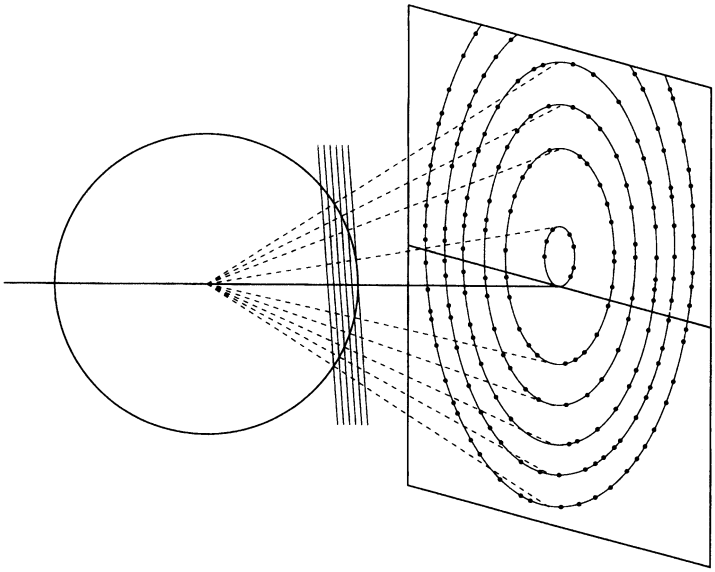


Fig. 4.5 The stack of reciprocal reflection planes that is nearly normal to the incident X-ray beam and intersecting the reflection sphere generates a set of concentric ellipses of reflections on the detector. (Reproduced with permission from Dauter and Wilson, 2001; International Union of Crystallography.)

In order to record a complete diffraction data set in the rotation technique the crystal is rotated or oscillated in incremental values $\Delta\varphi$ around the spindle axis for a given time, and the corresponding images are transferred to the controlling computer. Thereby, the whole asymmetric part of the reciprocal lattice must have passed through the Ewald sphere. One can imagine that the ellipses extend to lunes with a width proportional to the magnitude of $\Delta\varphi$. However, the lunes will overlap if $\Delta\varphi$ has been chosen too large, and consequently the choice of a suitable $\Delta\varphi$ is crucial for successful data collection. Another important point is that each reflection diffracts over a defined crystal rotation (referred to as the rocking curve or angular spread ξ), which is the combined effect of beam divergence δ and crystal mosaicity η. If we assume that ξ is less than $\Delta\varphi$, then some reflections will start and finish passing the Ewald sphere and hence diffract within one exposure. Their full intensity will be recorded on a single image, and these are called fully recorded reflections or "fullys". Reflections with ξ greater than $\Delta\varphi$ – which are referred to as partially recorded reflections or partials – will start reflecting on one exposure and end on the next. It is also possible that they extend over several adjacent exposures. We will see later how partially recorded reflections can be used in the post-refinement, scaling, and averaging of X-ray diffraction data.

As mentioned earlier, an X-ray data set will be complete if the whole asymmetric unit of the reciprocal lattice has passed the reflection sphere. The simplest way to achieve this is to rotate the crystal about 360° around the spindle axis. However, it

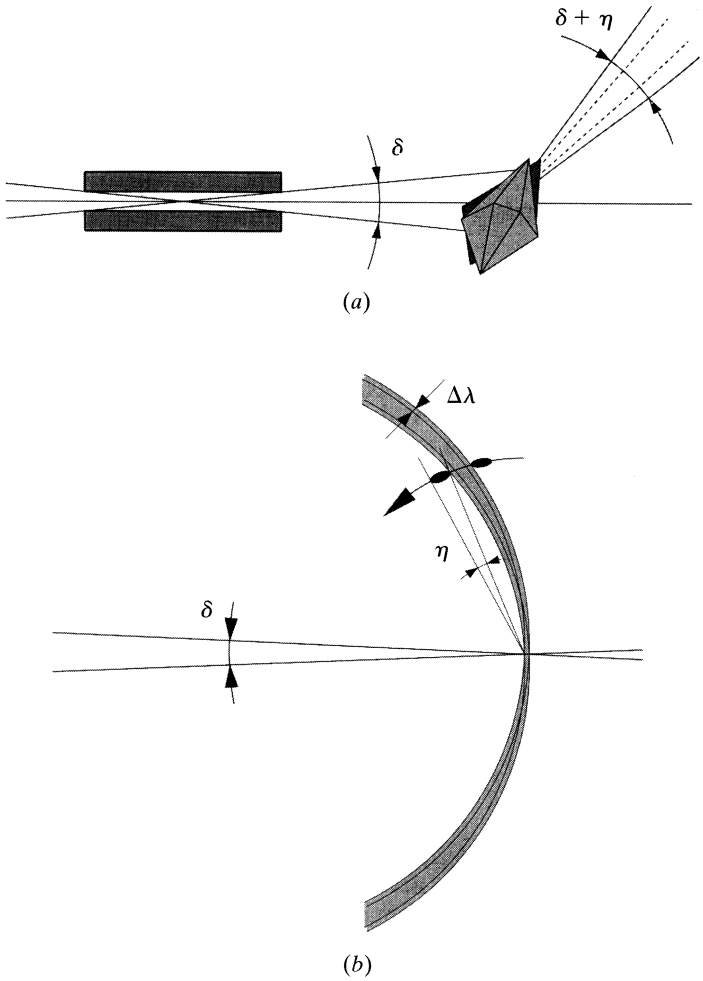


Fig. 4.6 Schematic representation of beam divergence δ and crystal mosaicity ε. (a) In direct space; (b) in reciprocal space, where the additional thickness of the reflection sphere results from the finite wavelength bandpass $\delta\lambda/\lambda$. (Reproduced with permission from Dauter and Wilson, 2001; International Union of Crystallography.)

is evident that this generates a highly redundant data set and a huge amount of data. Strategies to carry out an efficient data collection have been elaborated and are discussed by Dauter and Wilson (2001). The program MOSFLM offers an option STRATEGY, which proposes an optimal way of recording a data set with the highest attainable completeness and sufficient redundancy.

In this context, the number of measurable reflections is still of interest. They lie within the resolution sphere with radius $1/d_m$ and volume $(4/3)\pi/d_m^3$, where d_m was the maximal resolution of the diffraction data. This further implies that

d_m is greater than $\lambda/2$. Hence, we must divide the volume of the resolution sphere by the volume of the unit-cell of the reciprocal lattice $V^* = 1/V$, and take into account their centering to obtain the number of measurable reflections N according to

$$N = \frac{4}{3}\pi \frac{V}{d_m^3 n} \ , \tag{4.10 a}$$

where $n = 1$ for a primitive lattice, $n = 2$ for face- and body-centered lattices, and $n = 4$ for all face-centered lattices. For a primitive unit cell with a 100 Å cell constant in each direction and a resolution of 2 Å, about 500 000 reflections can be measured. For structures of large molecular assemblies (e.g., viral capsids), this may be about 10 million reflections, assuming a resolution of only 3 Å. This leads to the production of huge amounts of raw data that must be stored and processed.

4.5 Integration of Diffraction Spots

In order to obtain an optimal integration of the diffraction spots, a row of parameters must be determined either in advance or during the integration process. A prerequisite for autoindexing is knowledge of the wavelength λ, the crystal-to-detector distance, and the camera constants (CCX, CCY, CCOM). After autoindexing, the unit cell parameters should be refined with the highest possible accuracy of a few parts per thousand, and the missetting angles (PHIX, PHIY, PHIZ) should not deviate from zero by more than $\pm 0.05^\circ$. The crystal mosaicity can be estimated by visual inspection, and will be refined during the data evaluation process. MOSFLM refines a couple of additional parameters: YSCALE, the relative scale factor in the detector Y_d direction; and TILT and TWIST, deviations from normal incidence on the detector. TILT is a rotation about a horizontal axis, and TWIST about a vertical axis, while ROFF and TOFF are the radial and tangential offsets for Mar Research scanners. The refinement of DIST and missetting angles allows for crystal movement during data collection. Nonorthogonality of the incident X-ray beam and the rotation axis (if not allowed for), or an off-center crystal, will also give rise to apparent changes in crystal orientation with spindle axis rotation.

The integrated intensities can be determined using either of two different techniques, namely summation integration and profile fitting. Here, we describe the summation integration in greater detail. The 3D representation of the pixel raw data around a reflection is shown in Figure 4.7. The peak becomes apparent as an elevation from a background that is generated by X-ray scattering from air, the sample holder, and the sample itself. It is evident that a peak and the background area around the peak must be defined, and for this purpose a rectangular box of pixels (the measurement box) is centered on the raster coor-

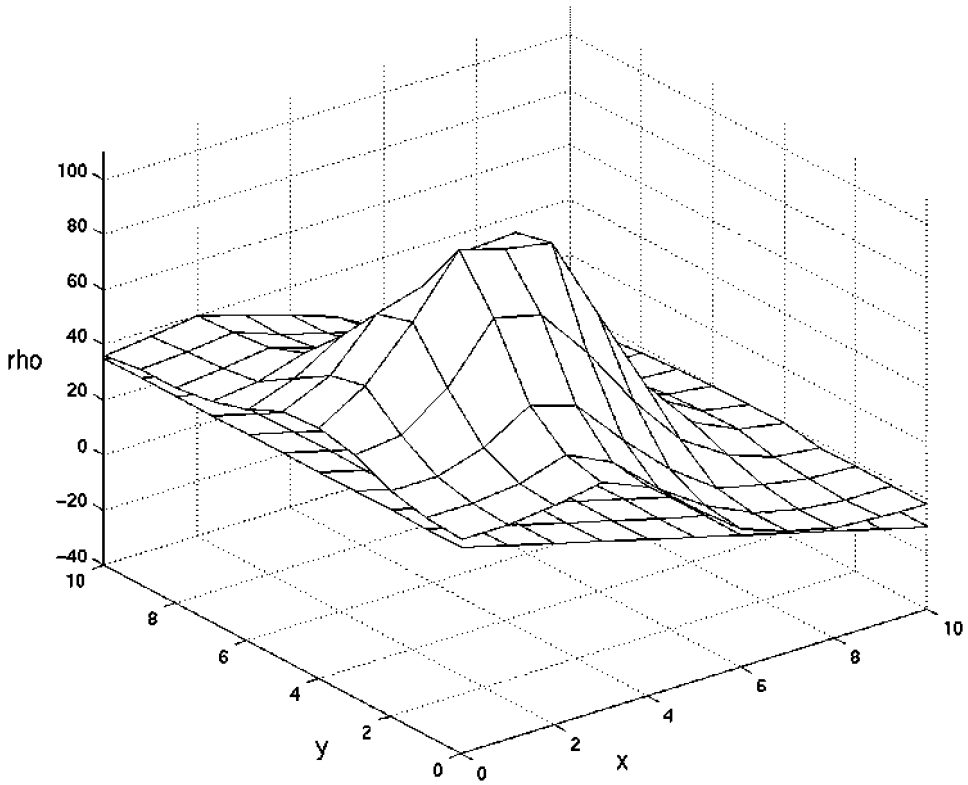


Fig. 4.7 3D representation of the pixel raw data around a reflection. (Figure was produced with MATCAB; Math Works, Inc., 2005.)

dinate nearest the calculated reflection position (Fig. 4.8). Each pixel within the box is classified as being either a background or peak pixel (or neither). This mask can be defined by the user, or the classification can be made automatically by the program. The background parameters NRX, NRY and NC can be optimized by maximizing the ratio of the intensity divided by its standard deviation.

The background can be reasonably represented by a plane over the area of a diffraction spot. With p_i, q_i, the pixel coordinates, the background total counts ρ_{bi} is given by $\rho_{bi} = ap + bq + c$, where a, b and c are constants. The background-subtracted total counts ρ of a pixel is obtained from $\rho_i - \rho_{bi}$ and the constants a, b and c defining the best background plane are determined by minimizing

$$R_1 = \sum_{i=1}^{n} w_i(\rho_i - a\rho_i - bq_i - c)^2 \,, \tag{4.10}$$

where ρ_i is the total counts at the pixel with coordinates p_i, q_i with respect to the center of the measurement box and the summation is over the n back-

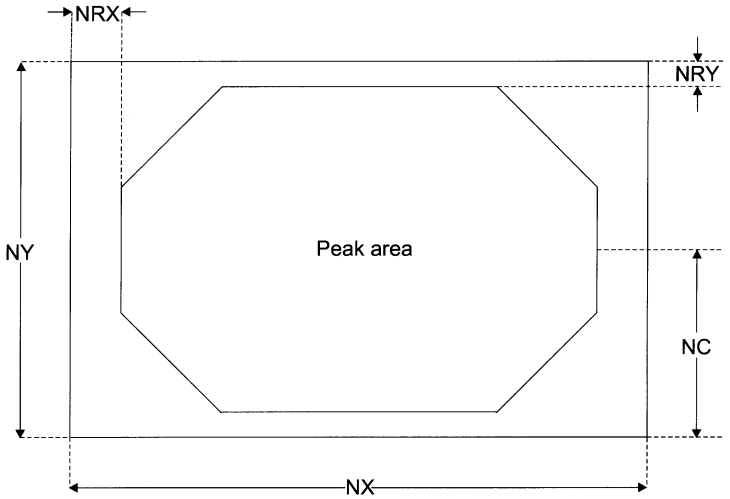


Fig. 4.8 The measurement-box definitions used in MOSFLM. The measurement box has overall dimensions of NX by NY pixels (both odd integers). The separation between peak and background pixels is defined by the widths of the background margins (NRX and NTY) and the corner cutoff (NC). (Adapted from Leslie, 2001.)

ground pixels. w_i is a weight which should ideally be the inverse of the variance of ρ_i. The summation integration intensity I_S is then obtained by

$$I_S = \sum_i^m (\rho_i - ap_i - bq_i - c) \,, \tag{4.11}$$

where the summation is over the m pixels in the peak region. As outlined by Leslie (2001), we obtain for the variance in I_S

$$\sigma_{I_s}^2 = G[I_S + I_{bg} + (m/n)/I_{bg}] \,, \tag{4.12}$$

with the background intensity I_{bg} and gain of detector G, which converts pixel counts to equivalent X-ray photons. Provided that the background and peak areas are correctly defined, summation integration provides a method for evaluating integrated intensities that is both robust and free from systematic error.

For the treatment of weak reflections, where the peak area contains very little signal from the Bragg intensity, profile fitting (Rossmann, 1979) delivers an improvement of the signal-to-noise ratio. In this approach, "standard" profiles are determined from well-recorded reflections for different areas of the detector and applied to the reflections. A detailed discussion of integration by profile fitting is given by Leslie (2001).

4.6
Post-Refinement, Scaling, and Averaging of Diffraction Data

As described in Section 4.5, a number of parameters must be refined in order to obtain an optimal integration of the diffraction data. In the past, this has been done by minimizing the sum of the differences between observed and calculated spot positions. However, it transpires that post-refinement procedures (Rossmann et al., 1979; Winkler et al., 1979), which make use of the estimated φ centroids of observed reflections, generally provide more accurate estimates. This is because spot positions are affected by residual spatial distorsions (after applying appropriate corrections) and estimated φ centroids do not obey a correlation between unit-cell parameters and crystal-to-detector distance, as is the case when using spot positions.

The objective in post-refinement is to compare the observed partiality p_{obs} with the calculated partiality p_{calc}. For this, one needs a reasonable model for p_{calc}. The effective mosaic spread m will give rise to a series of possible reflection spheres (Fig. 4.9). Their extreme positions will subtend an angle $2m$ at the origin of reciprocal space, and their centers will lie on a circle of radius $\delta = m/\lambda$. As the reciprocal lattice is rotated about the Oz axis, perpendicular to the mean direction of the X-ray beam Ox, a reciprocal lattice point P will gradually penetrate the effective thickness of the Ewald sphere.

Let q be a measure of the fraction of the path traveled by P between the extreme reflecting positions P_A and P_B. This path will be proportional to the fraction of the volume of a sphere, which has penetrated the reflection sphere, and corresponds to the fraction of the energy already diffracted. The volume of a sphere removed by a plane at a distance q from its surface is a good approximation for this. The volume p expressed as a fraction of the volume of the total sphere is then

$$p_{calc} = 3q^2 - 2q^3 \tag{4.13}$$

as q can be determined from the crystal setting parameters. We will not discuss the derivation of q from these parameters, which is provided by Rossmann et al. (1979). The observed partialities p_{obs} can be obtained from Eq. (4.14):

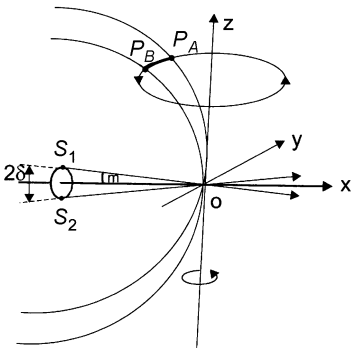


Fig. 4.9 Penetration of a reciprocal lattice point P into the Ewald sphere by rotation around Oz. The extremes of reflecting conditions at P_A and P_B are equivalent to X-rays passing along the lines S_1O and S_2O with centers of the Ewald spheres at S_1 and S_2 and subtending an angle of $2m$ at O. Therefore, in three dimensions, the extreme reflecting spheres will lie with their centers on a circle of radius $\delta = m/\lambda$ at $x = -1/\lambda$. (Adapted from Rossmann et al., 1979.)

$$p_{\text{obs}} = I_{him}/G_m I_{hi} \ , \tag{4.14}$$

where I_{hi} is the intensity of the ith measurement of reflection h, I_{him} the intensity contribution of reflection h_i recorded on frame m, and G_m the inverse linear scale factor for the respective frame m. Post-refinement is now done by minimizing the function

$$E = \sum w(p_{\text{obs}} - p_{\text{calc}})^2 \ , \tag{4.15}$$

where the sum is taken over all partial reflections for which one or more whole reflections are also observed, and w is a weight. Parameters, which can be refined, are the missetting angles, the independent unit-cell constants, the mosaic spread, and horizontal and vertical beam divergence. Details of the mathematical procedure can be taken from Rossmann et al. (1979).

There are many reasons why the different frames of a diffraction intensity data set must be scaled together to obtain the best estimates of intensities. The main causes are:

- radiation damage of the crystal, which leads to a weakening of the diffraction power and an increase of the mosaic spread, or to the death of the crystal, which requires the use of several crystals for a full data set;
- absorption effects of the crystal; and
- instabilities of the radiation source and the detector.

Nevertheless, a data set can be divided into different batches where the conditions were sufficiently constant to carry out scaling for the separate batches. Finally, the individual scaled batches will be scaled together. In the classical approach of Hamilton et al. (1965), only fully recorded reflections were used and the following function ψ was used for least-squares minimization

$$\psi = \sum_h \sum_i W_{hi}(I_{hi} - G_m I_h)^2 \ , \tag{4.16}$$

where I_h is the best estimate of the intensity of a reflection with reduced Miller indices h and W_{hi} is a weight for reflection h_i. For all unique reflections h, the values of I_h must correspond to a minimum in ψ. Thus,

$$\partial\psi/\partial I_h = 0 \ . \tag{4.17}$$

Carrying out this differentiation leads to the best least-squares estimate of a reflection

$$I_h = \sum_i W_{hi} G_m I_{hi} \sum_i W_{hi} G_m^2 \ . \tag{4.18}$$

Since ψ is a non-linear function of the scale factors G_m, the values of the scale factors must be determined by iterative non-linear least-squares techniques.

Recent advances in using frozen crystals of biological macromolecules had generated the situation that the individual images of many collected X-ray intensity data sets contain a high degree or only partially recorded reflections. This may be due to the increased mosaicity caused by the crystal freezing, the smaller size, and/or larger unit cell of used crystals accessible by the application of synchrotron radiation. An increase of the oscillation increment $\Delta\varphi$ is not feasible as it would introduce a high degree of overlapping reflections. Therefore, scaling techniques had to be developed that include partially measured intensities or even data sets consisting only of partially recorded intensities.

Provided that the reflection partiality p_{him} is known, the full intensity is estimated by

$$I_{hi} = I_{him}/p_{him}G_m\,. \tag{4.19}$$

Two methods are used for scaling including partially recorded reflections:

1. If a reflection h_i occurs on a number of adjacent frames and all parts of I_{him} are available in the data set, the generalized function ψ of Eq. (4.16) has the following form:

$$\psi = \sum_h \sum_i \sum_m \left\{ I_{him} - G_m \left[I_h - \sum_{m' \neq m} (I_{him'}/G_{m'}) \right] \right\}^2 . \tag{4.20}$$

The best least-squares estimate of I_h will be

$$I_h = \frac{\sum\limits_i \left[\sum\limits_m (I_{him}/G_m) \right] \left(\sum\limits_m W_{him} G_m^2 \right)}{\sum\limits_i \sum\limits_m G_m^2} = \frac{\sum\limits_i I_{hi} \sum\limits_m W_{him} G_m^2}{\sum\limits_i \sum\limits_m W_{him} G_m^2} \tag{4.21}$$

with

$$I_{hi} = \sum_m (I_{him}/G_m) \tag{4.22}$$

and m from 1 to the number of adjacent frames containing the full reflection.

2. If the theoretical partiality p_{him} of the partially recorded reflection h_{im} can be estimated, the generalized function ψ is derived as:

$$\psi = \sum_h \sum_i \sum_m W_{him} (I_{him} - G_m p_{him} I_h)^2 \tag{4.23}$$

and, using Eq. (4.19), the best least-squares estimate of I_h will then be

$$I_h = \frac{\sum_i \sum_m W_{him} G_m p_{him} I_{him}}{\sum_i \sum_m W_{him} G_m^2 p_{him}^2} \ . \tag{4.24}$$

The scale factor has been generalized to incorporate crystal damage (Otwinowski and Minor, 1997) in the form

$$G_{him} = G_m \exp\{-2B_m[\sin\theta_{hi}/\lambda]^2\} \ , \tag{4.25}$$

where B_m is a parameter describing the crystal disorder while frame m was recorded, θ_{hi} is the Bragg angle of reflection h_i, and λ is the X-ray wavelength.

When all scale factors have been determined they can be applied to the reflection intensities, and error estimates and the reflection intensities for the same reduced Miller indices can then be averaged. This is more complicated for method 2, and is performed in a two-step procedure.

Finally, estimates for the quality of data scaling and averaging are needed. Useful definitions of reliability factors R for scaled and averaged Bragg reflection intensities are:

$$R_{merge} = R_1 = \left[\left(\sum_h \sum_i |I_{hi} - \langle I_h \rangle| \right) \Big/ \sum_h \sum_i |I_{hi}| \right] \times 100\% \tag{4.26}$$

$$R_2 = \left\{ \left[\sum_h \sum_i (I_{hi} - \langle I_h \rangle)^2 \right] \Big/ \sum_h \sum_i I_{hi}^2 \right\} \times 100\% \tag{4.27}$$

and

$$R_w = \left\{ \left[\sum_h \sum_i W_{hi} (I_{hi} - \langle I_h \rangle)^2 \right] \Big/ \sum_h \sum_i W_{hi} I_{hi}^2 \right\} \times 100\% \ . \tag{4.28}$$

where R_{merge} is the linear, R_2 the square, and R_w the weighted R factor. A commonly used R factor is R_{merge}. This section is based on the report by van Beek et al. (2001), which discusses the subject in more detail.

References

Burzlaff, H., Zimmermann, H., de Wolff, P. M., Crystal lattices. In: Hahn, T. (Ed.), *International Tables for Crystallography*, Vol. A, Kluwer Academic Publishers, Dordrecht, **1992**.

Dauter, Z., Wilson, K. S., Principles of monochromatic data collection. In: Rossmann, M. G., Arnold, E. (Eds.), *International Tables for Crystallography*, Vol. F, pp. 177–195, Kluwer Academic Publishers, Dordrecht, **2001**.

Hamilton, W. C., Rollett, J. S., Sparks, R. A., *Acta Crystallogr.* **1965**, *18*, 129–130.

Howard, A. J., Nielsen, C., Xuong, Ng, H., *Methods Enzymol.* **1985**, *114*, 452–472.

Howard, A. J., Gilliland, G. I., Finzel, B. C., Poulos, T. L., Ohlendorf, D. H., Salemne, F. R., *J. Appl. Crystallogr.* **1987**, *20*, 383–387.

Leslie, A. G. W., *Acta Crystallogr.* **1999**, *D55*, 1696–1702.

Leslie, A. G. W., Integration of macromolecular diffraction data. In: Rossmann, M. G., Arnold, E. (Eds.), *International Tables for Crystallography*, Vol. F, pp. 212–217, Kluwer Academic Publishers, Dordrecht, **2001**.

Kabsch, W., *J. Appl. Crystallogr.* **1988a**, *21*, 67–81.

Kabsch, W., *J. Appl. Crystallogr.* **1988b**, *21*, 916–924.

Kabsch, W., Data collection and evaluation with program XDS. In: Sawyer, L., Isaac, N., Bailey, S. (Eds.), *Data Collection and Processing, Proceedings of the CCP4 Study Weekend, 29–30 January 1993*, SERC Daresbury Laboratory, Warrington, **1993**.

Messerschmidt, A., Pflugrath, J. W., *J. Appl. Crystallogr.* **1987**, *20*, 306–315.

Otwinowski, Z., Minor, W., *Methods Enzymol.* **1997**, *276*, 307–326.

Rossmann, M. G., *J. Appl. Crystallogr.* **1979**, *12*, 225–238.

Rossmann, M. G., Leslie, A. G. W., Abdel-Meguid, S. S., Tsukihara, T., *J. Appl. Crystallogr.* **1979**, *12*, 570–581.

Steller, I., Bolotowsky, R., Rossmann, M. G., *J. Appl. Crystallogr.* **1997**, *30*, 1036–1040.

Sparks, R. A., Trends in Minicomputer Hardware and Software, Part I. In: Ahmed, F. R., Huml, K., Sedlaček, B. (Eds.), *Crystallographic Computing Techniques*, pp. 452–467, Munksgaard, Copenhagen, **1976**.

Sparks, R. A., Data collection with diffractometers. In: Sayre, D. (Ed.), *Computational Crystallography*, pp. 1–18, Oxford University Press, New York, **1982**.

Winkler, F. K., Schutt, C. E., Harrison, S. C., *Acta Crystallogr.* **1979**, *A35*, 901–911.

Van Beek, C. G., Bolotkovsky, R., Rossmann, M. G., The use of partially recorded reflections for post refinement, scaling and averaging. In: Rossmann, M. G., Arnold, E. (Eds.), *International Tables for Crystallography*, Vol. F, pp. 236–242, Kluwer Academic Publishers, Dordrecht, **2001**.

5
Methods for Solving the Phase Problem

5.1
Isomorphous Replacement

5.1.1
Preparation of Heavy-Metal Derivatives

If one can attach one or several heavy-metal atoms at defined binding site(s) to the protein molecules without disturbing the crystalline order, one can use such isomorphous heavy-atom derivatives for the phase determination. This so-called method of isomorphous replacement was introduced by Perutz and co-workers in 1954 (Green et al., 1954). The lack of isomorphism can be monitored by a change in the unit cell parameters compared with the native crystal and a deterioration in the quality of the diffraction pattern. The preparation of heavy-atom derivatives is undertaken by soaking the crystals in mother liquor containing the dissolved heavy-metal compound. Soaking times may range from several minutes to months, while the concentrations of the heavy-metal compound may vary from tenths of millimolar to 50 mM.

The favorite heavy atoms to be used are Hg, Pt, U, Pb, Au, and the rare earth metals. Potential ligands can be classified as hard and soft ligands according to Pearson (1969). *Hard ligands* are electronegative and undergo electrostatic interactions. In proteins, such ligands are glutamate, aspartate, terminal carboxylates, hydroxyls of serines and threonines and in the buffer acetate, citrate and phosphate. By contrast, *soft ligands* are polarizable and form covalent bonds; they include as cysteine, cystine, methionine and histidine in proteins, and Cl^-, Br^-, I^-, S-ligands, CN^- and imidazole in the buffer solution.

Metals are classified according to their preference for hard or soft ligands. Class (a) metals, which bind preferentially to hard ligands, comprise the cations of A-metals such as alkali and alkaline earth metals, the lanthanides, some actinides and groups IIIA, IVA, VA and VIA of the transition metals.

IIIA	IVA	VA	VIA
Sc	Ti	V	Cr
Y	Zr	Nb	Mo
Lanth.	Hf	Ta	W

Class (b) metals are rather soft and polarizable, and can form covalent bonds to soft ligands; they include heavy metals at the end of the transition metal groups such as Hg, Pt and Au.

Their complexes are covalent and often anionic, as for example $Pt(CN)_4^{2-}$, $Au(CN)_2^{2-}$, $[PtCl_4]^{2-}$ and HgI_4^{2-}. The most stable ions are formed with the softest ligands, so that methionine, cysteine and imidazole will replace Cl^- in $[PtCl_4]^{2-}$ but not in $[Pt(CN)_4]^{2-}$. Class (b) metals are at the end of the transition groups and have outer shells of polarizable *d*-electrons.

There are also ions in the middle and towards the end of the first subgroups with transient properties between the class (a) and (b) metals. These can be ranked with increasing class (b) character:

$$Fe^{2+} < Co^{2+} < Ni^{2+} < Cu^{2+} < Zn^{2+} .$$

For example, Zn^{2+}, with the highest class (b) character, is usually found with soft sulfur and imidazole ligands, though carboxylate and water ligands are also found. There is also an influence of the precipitant, buffer, pH and metal salt concentration on binding of the metal. For example, the pH determines the protonation state of putative binding ligands, and with it the binding properties of the ligand.

In summary, in the protein, the class (b) metals Hg, Pt and Au and their complex compounds bind to soft ligands such as cysteine, histidine or methionine, while the class (a) metals U and Pb bind to hard ligands such as the carboxylate groups of glutamate or aspartate.

Information on the preparation and characterization of heavy-atom derivatives has been collected (Carvin et al., 1991; Islam et al., 1998), and this heavy-atom data bank is available at http://www.bmm.icnet.uk/had/. The data bank contains information on heavy-atom derivatives for approximately 1000 protein crystals.

5.1.2 Single Isomorphous Replacement

The structure factor $\mathbf{F}_{PH}$ for the heavy-atom derivative structure (Fig. 5.1) becomes (Eq. 5.1):

$$\mathbf{F}_{PH} = \mathbf{F}_P + \mathbf{F}_H \tag{5.1}$$

where $\mathbf{F}_P$ is the structure factor of the native protein and $\mathbf{F}_H$ is the contribution of the heavy atoms to the structure factor of the derivative. The isomorphous differences, F_{PH}–F_P, which can be calculated from experimental intensity data sets of the native and derivative protein, correspond to the distance CB in Figure 5.1. By inspection of Figure 5.1 we can derive this expression by using simple trigonometry. For example, we obtain for the distances BD, OC, OD: $BD = F_H \cos(a_{PH} - a_H)$, $OC = F_P$ and $OD = F_P \cos(a_P - a_{PH})$. Now, we can calculate

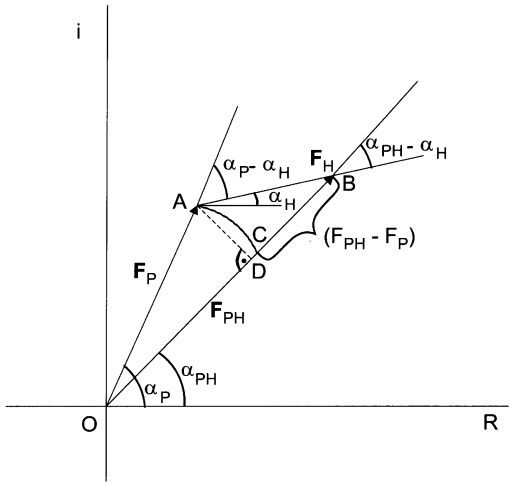


Fig. 5.1 Vector diagram for the vector addition of the structure factor of the native protein $\mathbf{F}_P$ and the heavy-atom contribution $\mathbf{F}_H$ to the heavy-atom derivative structure factor $\mathbf{F}_{PH}$. The relevant phase angles in Eq. (5.3) that are used for its derivation are also shown.

$$\mathrm{BE} = \mathrm{BD} - (\mathrm{OE\text{-}OD}) = F_{PH} - F_P \tag{5.2}$$

Then, we substitute the corresponding expressions and obtain

$$\begin{aligned} F_{PH} - F_P &= F_H \cos(a_{PH} - a_H) - F_P\{1 - \cos(a_P - a_{PH})\} \\ &= F_H \cos(a_{PH} - a_H) - 2F_P \sin^2\left(\frac{a_P - a_{PH}}{2}\right) \end{aligned} \tag{5.3}$$

If F_H is small compared with F_P and F_{PH}, the sine term will be very small and we have (Eq. 5.4)

$$F_{PH} - F_P \approx F_H \cos(a_{PH} - a_H) \tag{5.4}$$

When vectors $\mathbf{F}_P$ and $\mathbf{F}_H$ are colinear, then:

$$|\mathbf{F}_{PH} - \mathbf{F}_P| = F_H \tag{5.5}$$

The square of the isomorphous differences, F_{PH}–F_P, can be used as coefficients in a Patterson synthesis. Hence, we obtain

$$\begin{aligned} &(F_{PH} - F_P)^2 = 4F_P^2 \sin^4\left(\frac{a_P - a_{PH}}{2}\right) && \text{(i)} \\ &+F_H^2 \cos^2(a_{PH} - a_H) && \text{(ii)} \\ &-4F_P F_H \sin^2\left(\frac{a_P - a_{PH}}{2}\right) \times \cos(a_{PH} - a_H) && \text{(iii)} \end{aligned} \tag{5.6}$$

It is a theorem of Fourier theory that the Fourier transform of the sum of Fourier coefficients is equal to the sum of the Fourier transforms of the individual

Fourier coefficients. Here, there are three different terms. $(a_P - a_{PH})$ is small if F_H is small, and term (i), which gives the protein–protein interaction, will be of low weight. The transform of term (iii) is zero if sufficient terms are included. However, if $F_H \ll F_P$, $(a_P - a_{PH})$ is effectively random and term (ii) will give heavy atom vectors with half the expected peak heights (Eq. 5.7):

$$F_H^2 \cos^2(a_{PH} - a_H) = \frac{1}{2} F_H^2 + \frac{1}{2} F_H^2 \cos 2(a_{PH} - a_H) \tag{5.7}$$

with the second term on the right contributing only noise to the Patterson map because the angles a_{PH} and a_H are not correlated. Such an isomorphous heavy-atom difference Patterson map allows determination of the positions of the heavy metals on the condition of isomorphism and a not too-large heavy-atom partial structure.

In this context we introduce the structure factor for a centro-symmetric structure, and explain the meaning of centric zones. In a centro-symmetric atomic structure we have for each atom with coordinates x_j, y_j, z_j a counterpart with coordinates $-x_j, -y_j, -z_j$. That means we must write the structure factor for this case as

$$\begin{aligned} F(hkl) &= \sum_{j=1}^{N/2} f_j \exp 2\pi i(hx_j + ky_j + lz_j) + \sum_{j=N/2+1}^{N} f_j \exp 2\pi i(-hx_j - ky_j - lz_j) \\ &= \sum_{j=1}^{N/2} f_j \{\cos 2\pi(hx_j + ky_j + lz_j + i \sin 2\pi(hx_j + ky_j + lz_j)\} \\ &\quad + \sum_{j=N/2+1}^{N} f_j \{\cos 2\pi(hx_j + ky_j + lz_j) - i \sin 2\pi(hx_j + ky_j + lz_j)\} \\ &= \sum_{j=1}^{N/2} 2f_j \cos 2\pi(hx_j + ky_j + lz_j) \end{aligned} \tag{5.8}$$

This means that the structure factor for a centro-symmetric atomic structure is a real number and is either plus or minus only. Once the sign of the structure factor has been correctly determined, the error of its phase is zero. It was mentioned earlier that crystals of biological macromolecules exhibit acentric symmetries only. However, special reflection groups in a part of the allowed acentric space groups have real structure factors. These groups are denoted as centric zones, and their meaning is explained in the example of space group $P2_1$ (Fig. 5.2). In this space group, we have for each coordinate triple x_j, y_j, z_j a symmetry mate at $-x_j, y_j + 1/2, -z_j$ due to the 2_1-screw axis parallel **b**. For reflections of the (h 0 l) zone the structure factor becomes

$$F(hkl) = \sum_{j=1}^{N} f_j \exp 2\pi i(hx_j + 0y_j + lz_j) = \sum_{j=1}^{N} \exp 2\pi i(hx_j + lz_j)\ . \tag{5.9}$$

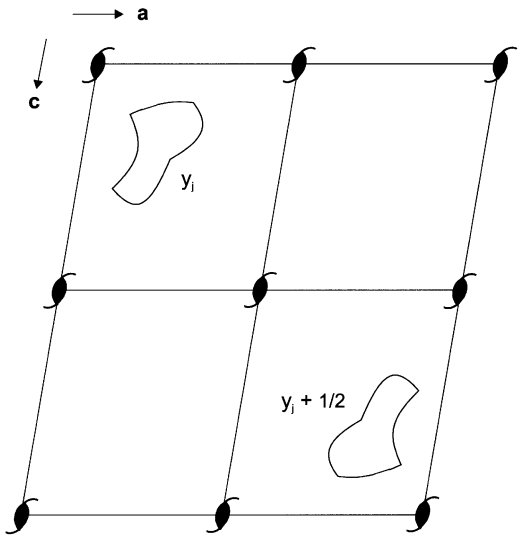


Fig. 5.2 Explanation of a centric zone in space group $P2_1$. Neglecting the y_j-coordinate leads to a projection of the atomic structure onto the **a**, **c**-plane. This projection is centrosymmetric.

The structure factor is independent of the y_j-coordinate and centro-symmetrical with respect to the x_j, z_j-coordinates. Therefore, the phases of reflections (h 0 l) are either plus or minus only. There is a simple rule for the occurrence of centric zones: they are perpendicular to even-numbered rotation and screw axes (regardless of the translational component).

It is important to know what intensity changes are generated by the attachment of heavy atoms to the macromolecule. According to Crick and Magdoff (1956), the relative root mean square intensity change is given by Eq. (5.10) for centric reflections:

$$\frac{\sqrt{\overline{(\Delta I)^2}}}{\overline{I_P}} = 2 \times \sqrt{\frac{\overline{I_H}}{\overline{I_P}}} \tag{5.10}$$

and by Eq. (5.11) for acentric reflections:

$$\frac{\sqrt{\overline{(\Delta I)^2}}}{\overline{I_P}} = \sqrt{2} \times \sqrt{\frac{\overline{I_H}}{\overline{I_P}}} \tag{5.11}$$

where $\overline{I_H}$ is the average intensity of the reflections if the unit cell were to contain the heavy atoms only, and $\overline{I_P}$ is the average intensity of the reflections of the native protein. Attaching one mercury atom (Z=80) to a macromolecule with varying molecular mass, and assuming 100% occupancy, gives the following average relative changes in intensity: 0.51 for 14 000 Da, 0.25 for 56 000 Da, 0.18 for 112 000 Da, 0.13 for 224 000 Da, and 0.09 for 448 000 Da. From this estimation it is evident that, with increasing molecular mass, more heavy atoms (or

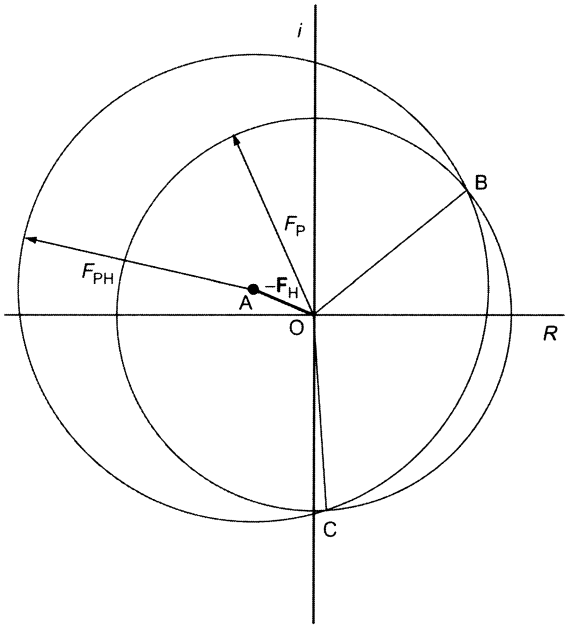


Fig. 5.3 Harker construction for the phase calculation by the method of single isomorphous replacement.

for large molecular masses heavy-metal clusters such as $Ta_6Br_{12}^{2+}$ (Knäblein et al., 1997) must be introduced to generate intensity changes which can be statistically measured (precision for intensity measurements between 5 and 10%) and which are sufficient for the phasing.

The phase calculation for single isomorphous replacement can be seen from the so-called Harker construction for this case (Fig. 5.3). $\mathbf{F}_H$, which can be calculated from the known heavy-atom positions, is drawn in its negative direction from the origin O ending at point A. Circles are drawn with radii F_P and F_{PH} from points O and A, respectively. The connections of the intersection points of both circles B and C with origin O determine two possible phases for $\mathbf{F}_P$. This means that the single isomorphous replacement leaves an ambiguity in the phase determination for the acentric reflections.

5.1.3 Multiple Isomorphous Replacement

The phase ambiguity can be overcome if two or more isomorphous heavy-atom derivatives are used which exhibit different heavy-atom partial structures. In Figure 5.4 the Harker construction for two different heavy-atom derivatives is shown. In addition to Figure 5.3, $-\mathbf{F}_{H2}$ is drawn from the origin O and a third circle with radius F_{PH2} is inserted around its end-point B. The intersection

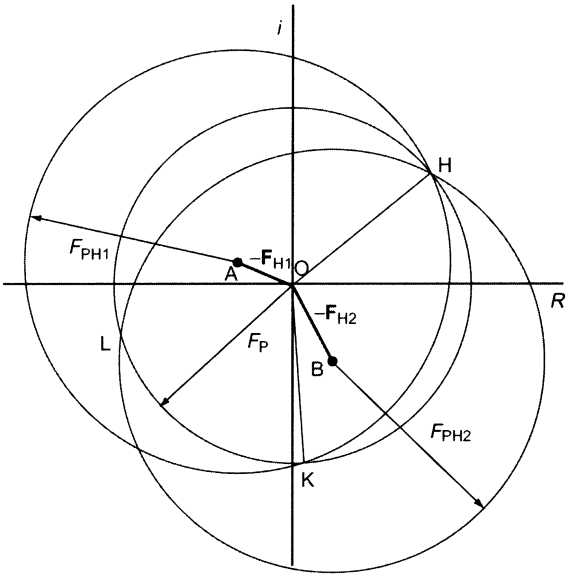


Fig. 5.4 Harker construction for the phase calculation by the method of MIR for two different heavy-atom derivatives PH1 and PH2.

point, H, of all three circles determines the protein phase, a_P. In the case of n isomorphous derivatives there are $n + 1$ circles which have one common intersection point whose connection to origin O determines the protein phase, a_P.

5.2 Anomalous Scattering

5.2.1 Theoretical Background

So far, in the normal Thomson scattering of X-rays, the electrons in the atom have been treated as free electrons that vibrate as a dipole-oscillator in response to the incident electromagnetic radiation and generate elastic scattering of the X-rays. However, the electrons are bound to atomic orbitals in atoms, and this treatment is valid only if the frequency ω of the incident radiation is large compared with any natural absorption frequency ω_{kn} of the scattering atom. For the light atoms in biological macromolecules (H, C, N, O, S, P) with frequency ω of the used radiation (in the range of 0.4 to 3.5 Å), this condition is fulfilled and these atoms really scatter normally. For heavier elements, the assumption $\omega >> \omega_{kn}$ is no longer valid, and the frequency ω may be higher for some and lower for other absorption frequencies. If ω is equal to an absorption frequency ω_{kn}, then absorption of radia-

tion will occur which is manifested by the ejection of a photoelectron with an energy corresponding to the ionization energy for this electron. This transition goes to a state in the continuous region because the discrete energy states are all occupied in the atom. The absorption frequencies for the K, L, or M shells are connected with the corresponding absorption edges, which are characterized by a sharp drop in the absorption curve (absorption versus λ) at the edge position. It is evident that the scattering from those electrons with their resonance frequencies close or equal to the frequency of the incident radiation will deliver a special contribution, which is called "anomalous scattering".

The classical treatment (see James, 1960) is briefly outlined here. It is assumed that the atoms scatter as if they contain electric dipole-oscillators having certain definite natural frequencies. The classical differential equation of the motion of a particle of mass m and charge e in an alternating electric field $\mathbf{E} = \mathbf{E}_0 e^{i\omega t}$ is:

$$\ddot{\mathbf{x}} + k\dot{\mathbf{x}} + \omega_s^2 \mathbf{x} = \frac{e\mathbf{E}_0}{m} e^{i\omega t} \tag{5.12}$$

where the damping factor, k, is proportional to the velocity of the displayed charge and ω_s is the natural circular frequency of the dipole if the charge is displaced. The steady-state solution for this equation for the moment of the dipole which executes forced oscillations of frequency under the action of the incident wave is:

$$\mathbf{M} = e\mathbf{x} = \frac{e^2}{m} \frac{\mathbf{E}_0 e^{i\omega t}}{\omega_s^2 - \omega^2 + ik\omega} \tag{5.13}$$

The amplitude A of the scattered wave at unit distance in the equatorial plane is given by:

$$A = \frac{e^2}{mc^2} \frac{\omega^2 E_0}{\omega_s^2 - \omega^2 + ik\omega} \tag{5.14}$$

The scattering factor of the dipole, f, is now defined as the ratio of the amplitude scattered by the oscillator to that scattered by a free classical electron under the same conditions. This amplitude at unit distance and in the equatorial plane is given by:

$$A' = -\frac{e^2}{mc^2} E_0 \tag{5.15}$$

Hence, we obtain Eq. (5.16) for f:

$$f = \frac{\omega^2}{\omega^2 - \omega_s^2 - ik\omega} \tag{5.16}$$

If f is positive the scattered wave has a phase difference of π with respect to the primary beam (introduced by the negative sign in the equation for A'). If $\omega >>$

ω_s, f is unity. In the case of $\omega << \omega_s$, f is negative and the dipole then scatters a wave in phase with the primary beam.

Equation (5.16) can be split into real and imaginary parts, so that we obtain Eq. (5.17):

$$f = f' + if'' \tag{5.17}$$

with Eqs. (5.18) and (5.19):

$$f' = \frac{\omega^2(\omega^2 - \omega_s^2)}{(\omega^2 - \omega_s^2)^2 + k^2\omega^2} \tag{5.18}$$

$$f'' = \frac{K\omega^3}{(\omega^2 - \omega_s^2)^2 + k^2\omega^2} \tag{5.19}$$

We now extend this for an atom consisting of s electrons each acting as a dipole-oscillator with oscillator strength $g(s)$ and resonance frequency ω_s. We have to multiply the contribution for each electron by $g(s)$ and form the sum over all electrons. For the total real part of the atomic scattering factor we obtain:

$$f' = \sum_s \frac{g(s)\omega^2}{\omega^2 - \omega_s^2} \tag{5.20}$$

which assumes that ω is not very nearly equal to ω_s, and a small damping. f' can be written as:

$$f' = f_0 + \Delta f' = \sum_s g(s) + \sum_s \frac{g(s)\omega_s^2}{\omega^2 - \omega_s^2} \tag{5.21}$$

For free electrons, we have $\omega_s = 0$ and $f' = f_0 = \sum_s g(s)$. The real part of the increment of the scattering factor is due to the binding of electrons. $\Delta f'$ is the dispersion component of the anomalous scattering.

If ω is comparable to ω_s but slightly greater, $ik\omega$ must not be neglected. f becomes complex:

$$f = f' + if'' = f_0 + \Delta f' + i\Delta f'' \tag{5.22}$$

The imaginary part lags $\pi/2$ behind the primary wave – that is, it is always $\pi/2$ in front of the scattered wave. $\Delta f''$ is known as the absorption component of the anomalous scattering.

In the quantum mechanical treatment of the problem the oscillator strengths are calculated from the atomic wave functions. Hönl (1933), in a series of theoretical investigations, used hydrogen-like atomic wave functions. In the frame of this approach, to each natural dipole frequency ω_s in the classical expression there corresponds in the quantum expression a frequency ω_{kn}, which is the

Bohr frequency associated with transition of the atom from the energy state k to the state n, in which it is supposed to remain during the scattering. Modern quantum mechanical calculations of anomalous scattering factors on isolated atoms, based on relativistic Dirac–Slater wave functions, have been carried out by Cromer and Liberman (1970). It follows from the theory of the anomalous scattering of X-rays that f_0 is real and independent of the wavelength of the incident X-rays, but dependent on the scattering angle.

$\Delta f'$ and $\Delta f''$ depend on the wavelength, λ, of the incident radiation, but are virtually independent of the scattering angle.

5.2.2 Experimental Determination

$\Delta f''$ is related to the atomic absorption coefficient μ_0 by Eq. (5.23):

$$\Delta f''(\omega) = \frac{mc\omega}{4\pi e^2}\mu_0(\omega) \tag{5.23}$$

$\Delta f'$ can now be calculated by the Kramers–Kronig transformation:

$$\Delta f'(\omega) = \frac{2}{\pi}\int_0^\infty \frac{\omega'\Delta f''(\omega')}{\omega^2 - \omega'^2}\,d\omega' \tag{5.24}$$

As fluorescence is closely related to absorption, fluorescence measurements varying the X-ray radiation frequency are used to determine the frequency dependence of the dispersive components of the different chemical elements. Instead of the radiation frequency ω, the radiation is often characterized by its wavelength, λ, or photon energy, E. The dispersion correction terms $\Delta f'$ and $\Delta f''$ are often simply denoted f' and f''. Figure 5.5 shows the anomalous scattering factors near the absorption K edge of selenium from a crystal of *Escherichia coli* selenomethionyl thioredoxin. The spectrum was measured with tunable synchrotron radiation. Apart from the "white line" feature at the absorption edge, f'' drops by about 4 electrons approaching the edge from the short wavelength side; $\Delta f'$ exhibits a symmetrical drop of –8 electrons around the edge. Similar values can be observed at the K edges for Fe, Cu, Zn, and Br, whose wavelengths all lie in the range 0.9 to 1.8 Å, which is well suited to biological macromolecular X-ray diffraction experiments. For other interesting heavy atoms such as Sm, Ho, Yb, W, Os, Pt, and Hg the LII (Sm) or LIII edges are in this range. Here, the effects are even greater. Considerably larger changes are found for several lanthanides, such as Yb, where the minimum f' is –33 electrons and the maximum f'' is 35 electrons.

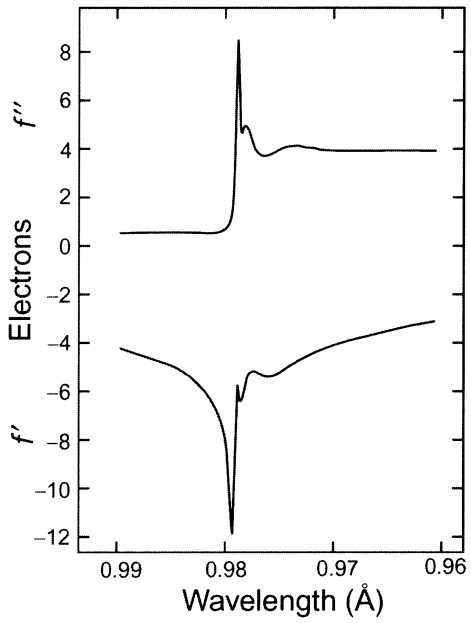


Fig. 5.5 Anomalous scattering factors near the absorption K-edge of selenium from a crystal by *E. coli* selenomethionyl thioredoxin. (Reproduced by permission of Academic Press, Inc., from Hendrickson and Ogata, 1997.)

5.2.3
Breakdown of Friedel's Law

Under the assumption that the crystal contains a group of anomalous scatterers, one can separate the contributions from the distinctive components of the scattering factor according to Hendrickson and Ogata (1997) to obtain:

$$^{\lambda}\mathbf{F}(\mathbf{h}) = {}^{\circ}\mathbf{F}_{\mathrm{N}}(\mathbf{h}) + {}^{\circ}\mathbf{F}_{\mathrm{A}}(\mathbf{h}) + {}^{\lambda}\mathbf{F}'_{\mathrm{A}}(\mathbf{h}) + i^{\lambda}\mathbf{F}''_{\mathrm{A}}(\mathbf{h}) \tag{5.25}$$

where ${}^{\circ}\mathbf{F}_{\mathrm{N}}$ is the contribution of the normal scatterers and ${}^{\circ}\mathbf{F}_{\mathrm{A}}$, ${}^{\lambda}\mathbf{F}'_{\mathrm{A}}$ and ${}^{\lambda}\mathbf{F}''_{\mathrm{A}}$ are the contributions for the corresponding components of the complex atomic form factor. For the centrosymmetric reflection, we obtain Eq. (5.26):

$$^{\lambda}\mathbf{F}(-\mathbf{h}) = {}^{\circ}\mathbf{F}_{\mathrm{N}}(-\mathbf{h}) + {}^{\circ}\mathbf{F}_{\mathrm{A}}(-\mathbf{h}) + {}^{\lambda}\mathbf{F}'_{\mathrm{A}}(-\mathbf{h}) + i^{\lambda}\mathbf{F}''_{\mathrm{A}}(-\mathbf{h}) \tag{5.26}$$

The geometric presentation for both structure factors is given in Figure 5.6. Inversion of the sign of $\mathbf{h}$ causes a negative phase angle for all contributions where the components of the scattering factor are real. For the f''-dependent part this is also valid but, owing to the imaginary factor i, this vector has to be constructed with a phase angle $+\pi/2$ with respect to ${}^{\circ}\mathbf{F}_{\mathrm{A}}(-\mathbf{h})$ and ${}^{\lambda}\mathbf{F}'_{\mathrm{A}}(-\mathbf{h})$. The resultant absolute values for ${}^{\lambda}\mathbf{F}(\mathbf{h})$ and ${}^{\lambda}\mathbf{F}(-\mathbf{h})$ are no longer equal, which means that their intensities (square of the amplitude) are different (breakdown of Friedel's law).

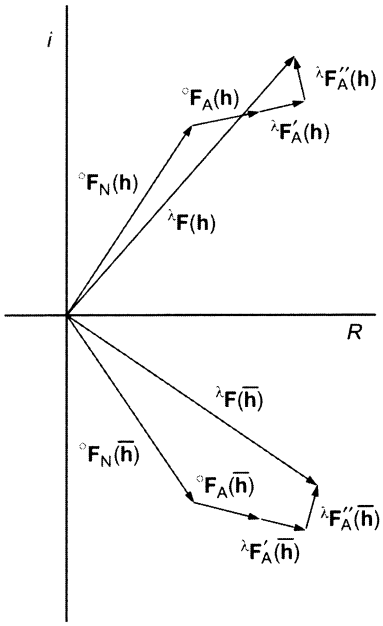


Fig. 5.6 Vector diagram explaining the breakdown of Friedel's law.

5.2.4
Anomalous Difference Patterson Map

One can show that (Eq. 5.27)

$$^{\lambda}F(\mathbf{h}) - {}^{\lambda}F(-\mathbf{h}) \approx \frac{2}{k}[{}^{\circ}F_{\mathrm{A}}(\mathbf{h}) + {}^{\lambda}F'_{\mathrm{A}}(\mathbf{h})]\sin(a_{\mathbf{h}} - a_{\mathrm{A}}) \tag{5.27}$$

where $a_{\mathbf{h}}$ is the phase angle of $^{\lambda}\mathbf{F}(\mathbf{h})$, a_{A} the phase angle of the anomalous scatterers and (Eq. 5.28)

$$k = \frac{^{\circ}F_{\mathrm{A}}(\mathbf{h}) + {}^{\lambda}F'_{\mathrm{A}}(\mathbf{h})}{^{\lambda}F''_{\mathrm{A}}(\mathbf{h})} \tag{5.28}$$

As coefficients for an anomalous difference Patterson, we obtain:

$$\Delta F^2_{\mathrm{ano}} = [{}^{\lambda}F(\mathbf{h}) - {}^{\lambda}F(-\mathbf{h})]^2 \sim \frac{4}{k^2}[{}^{\circ}F_{\mathrm{A}}(\mathbf{h}) + {}^{\lambda}F'_{\mathrm{A}}(\mathbf{h})]^2 \sin^2(a_{\mathbf{h}} - a_{\mathrm{A}})\ . \tag{5.29}$$

The ΔF_{ano}s will be maximal if the phase angle a_{A} is perpendicular to the phase angle $a_{\mathbf{h}}$, and zero if both vectors are colinear, which is opposite to the MIR case. The anomalous Patterson map contains peaks of the anomalous scatterers with heights proportional to half of $(4/k^2)[^{\circ}F_{\mathrm{A}}(\mathbf{h}) + {}^{\lambda}F'_{\mathrm{A}}(\mathbf{h})]^2$ owing to the $\sin^2$ term, and is therefore suited to determine the structure of the anomalous scatterers.

5.2.5
Phasing Including Anomalous Scattering Information

The combination of anomalous scattering information with isomorphous replacement permits the unequivocal determination of the protein phases, as shown in Figure 5.7. Using the anomalous scattering information alone gives two possible solutions for the protein phase characterized by the intersection points H and L in Figure 5.7. Combining it with the corresponding intensities from the native protein without the anomalous scatterers leaves only one solution for the protein phase (vector O – H in Fig. 5.7). The case in Figure 5.7 is called single isomorphous replacement anomalous scattering (SIRAS). Having *n* isomorphous heavy-atom derivatives, each with anomalous scattering contributions, the Harker construction can be extended for this situation and the phasing method is then designated multiple isomorphous replacement anomalous scattering (MIRAS).

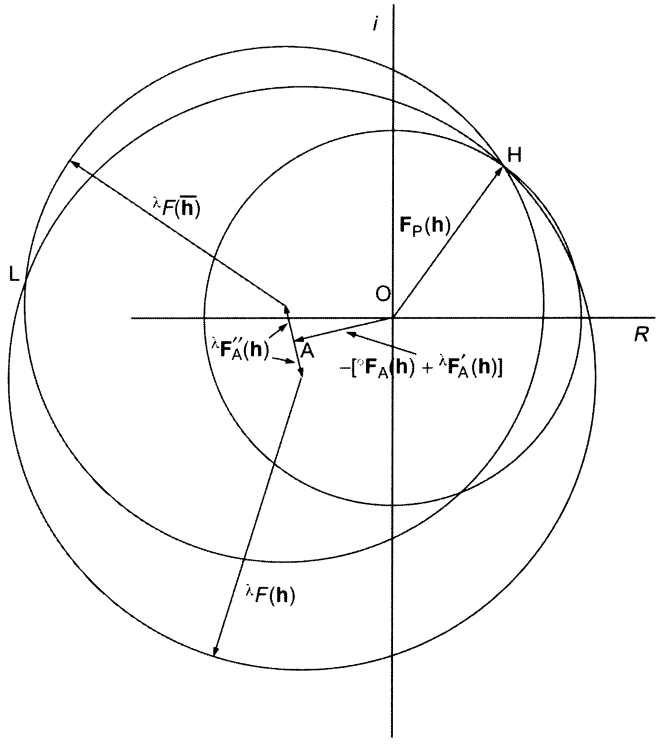


Fig. 5.7 Harker construction illustrating the phase determination combining information from anomalous scattering and isomorphous replacement.

5.2.6
The Multiwavelength Anomalous Diffraction (MAD) Technique

During the past few years, the MAD technique has matured to be a routine method, and has led to a revolution in biological macromolecular crystallography. If there are one or a few anomalous scatterers in the biological macromolecule it is possible to determine the whole spatial structure from one crystal (exact isomorphism) by the MAD technique. The anomalous scatterers may be either intrinsic, as in metalloproteins (e.g., Fe, Zn, Cu, Mo, Mn), or exogenous (e.g., Hg in a heavy-atom derivative or Se in selenomethionyl proteins; Hendrickson et al., 1990; Yang et al., 1990). A prerequisite for the MAD technique is well-diffracting crystals (resolution better than 2.8 Å), because the anomalous components of the atomic form factor are virtually independent of the diffraction angle and acquire increasing weight with increasing scattering angle. This advantageous property, together with exact isomorphism, serves for the determination of good phases down to the full resolution, and also leads to the production of excellent experimental MAD-phased electron density maps. A typical MAD experiment is carried out at three different wavelengths (tunable synchrotron radiation), at minimum f' and maximum f'' at the absorption edge of the anomalous scatterer(s) and at a remote wavelength where anomalous scattering effects are small.

The basic equations for the MAD technique as formulated by Hendrickson and Ogata (1997) are as follows. Equation (5.25) can be written as Eq. (5.30):

$$^{\lambda}\mathbf{F}(\mathbf{h}) = {}^{\circ}\mathbf{F}_{\mathrm{T}}(\mathbf{h}) + {}^{\lambda}\mathbf{F}_{\mathrm{A}}(\mathbf{h}) + i^{\lambda}\mathbf{F}''_{\mathrm{A}}(\mathbf{h}) \tag{5.30}$$

where

$$^{\circ}\mathbf{F}_{\mathrm{T}} = {}^{\circ}\mathbf{F}_{\mathrm{N}} + {}^{\circ}\mathbf{F}_{\mathrm{A}} \tag{5.31}$$

with subscript T for the totality of atoms in the structure.

Furthermore, we have Eqs. (5.32)–(5.35):

$$^{\circ}\mathbf{F}_{\mathrm{T}}(f^{\circ}) = {}^{\circ}F_{\mathrm{T}} \exp(i^{\circ}\phi_{\mathrm{T}}) \tag{5.32}$$

$$^{\circ}\mathbf{F}_{\mathrm{A}}(f^{\circ}) = {}^{\circ}F_{\mathrm{A}} \exp(i^{\circ}\phi_{\mathrm{A}}) \tag{5.33}$$

$$^{\lambda}\mathbf{F}'_{\mathrm{A}} = f(f') \tag{5.34}$$

$$^{\lambda}\mathbf{F}''_{\mathrm{A}} = f(f'') \tag{5.35}$$

In the common case of a single kind of anomalous scatterer, we obtain Eqs. (5.36) and (5.37):

$$^{\lambda}\mathbf{F}'_{\mathrm{A}} = \frac{f'(\lambda)}{f^{\circ}}\,{}^{\circ}\mathbf{F}_{\mathrm{A}} \tag{5.36}$$

$$ {}^{\lambda}\mathbf{F}''_{\mathrm{A}} = \frac{f''(\lambda)}{f^{\circ}} {}^{\circ}\mathbf{F}_{\mathrm{A}} \tag{5.37} $$

Separating the experimentally observable squared amplitude into wavelength-dependent and wavelength-independent terms gives Eq. (5.38):

$$ \begin{aligned} {}^{\lambda}F(\pm\mathbf{h})^2 = {}&{}^{\circ}F_{\mathrm{T}}^2 + a(\lambda){}^{\circ}F_{\mathrm{A}}^2 + b(\lambda){}^{\circ}F_{\mathrm{T}}{}^{\circ}F_{\mathrm{A}}\cos({}^{\circ}\phi_{\mathrm{T}} - {}^{\circ}\phi_{\mathrm{A}}) \\ &\pm c(\lambda){}^{\circ}F_{\mathrm{T}}{}^{\circ}F_{\mathrm{A}}\sin({}^{\circ}\phi_{\mathrm{T}} - {}^{\circ}\phi_{\mathrm{A}}) \end{aligned} \tag{5.38} $$

with (Eqs. 5.39–5.41)

$$ a(\lambda) = \frac{f'^2 + f''^2}{f^{\circ 2}} \tag{5.39} $$

$$ b(\lambda) = 2\frac{f'}{f^{\circ}} \tag{5.40} $$

$$ c(\lambda) = 2\frac{f''}{f^{\circ}} \tag{5.41} $$

The derivation of the formula for ${}^{\lambda}F(\mathbf{h})^2$ is illustrated in detail:

${}^{\lambda}F(\mathbf{h})^2$ is obtained from the triangle formed by the vectors ${}^{\lambda}\mathbf{F}(\mathbf{h})$, ${}^{\circ}\mathbf{F}_{\mathrm{T}}(\mathbf{h})$ and $\mathbf{a}$ (Fig. 5.8) by use of the cosine rule. The absolute values of the vectors are represented in italics, and the relevant angle is $(180^{\circ} - {}^{\circ}\phi_{\mathrm{A}} - \delta)$. Hence, we obtain:

$$ \begin{aligned} {}^{\lambda}F(\mathbf{h})^2 = {}&{}^{\circ}F_{\mathrm{T}}^2 + \left(\frac{f'^2 + f''^2}{f^{\circ 2}}\right){}^{\circ}F_{\mathrm{A}}^2 - 2 \times {}^{\circ}F_{\mathrm{T}} \times \left(\frac{f'^2 + f''^2}{f^{\circ 2}}\right)^{1/2} \\ &\times {}^{\circ}F_{\mathrm{A}} \times \cos(\pi - \delta - {}^{\circ}\phi_{\mathrm{A}} + {}^{\circ}\phi_{\mathrm{T}}) \end{aligned} \tag{5.42} $$

The cosine term in Eq. (5.42) can be obtained using some basic trigonometry:

$$ \cos(\pi + ({}^{\circ}\phi_{\mathrm{T}} - {}^{\circ}\phi_{\mathrm{A}} - \delta)) = -\cos({}^{\circ}\phi_{\mathrm{T}} - {}^{\circ}\phi_{\mathrm{A}} - \delta) \tag{5.43} $$

$$ \cos(({}^{\circ}\phi_{\mathrm{T}} - {}^{\circ}\phi_{\mathrm{A}}) - \delta) = \cos({}^{\circ}\phi_{\mathrm{T}} - {}^{\circ}\phi_{\mathrm{A}}) \times \cos\delta + \sin({}^{\circ}\phi_{\mathrm{T}} - {}^{\circ}\phi_{\mathrm{A}}) \times \sin\delta \tag{5.44} $$

with

$$ \cos\delta = \left(\frac{f'}{f^{\circ}}\right) \times {}^{\circ}F_{\mathrm{A}} \Bigg/ \left(\frac{f'^2 + f''^2}{f^{\circ 2}}\right)^{1/2} \times {}^{\circ}F_{\mathrm{A}} = \left(\frac{f'}{f^{\circ}}\right) \Bigg/ \left(\frac{f'^2 + f''^2}{f^{\circ 2}}\right)^{1/2} \tag{5.45} $$

$$ \sin\delta = \left(\frac{f''}{f^{\circ}}\right) \times {}^{\circ}F_{\mathrm{A}} \Bigg/ \left(\frac{f'^2 + f''^2}{f^{\circ 2}}\right)^{1/2} \times {}^{\circ}F_{\mathrm{A}} = \left(\frac{f''}{f^{\circ}}\right) \Bigg/ \left(\frac{f'^2 + f''^2}{f^{\circ 2}}\right)^{1/2} \tag{5.46} $$

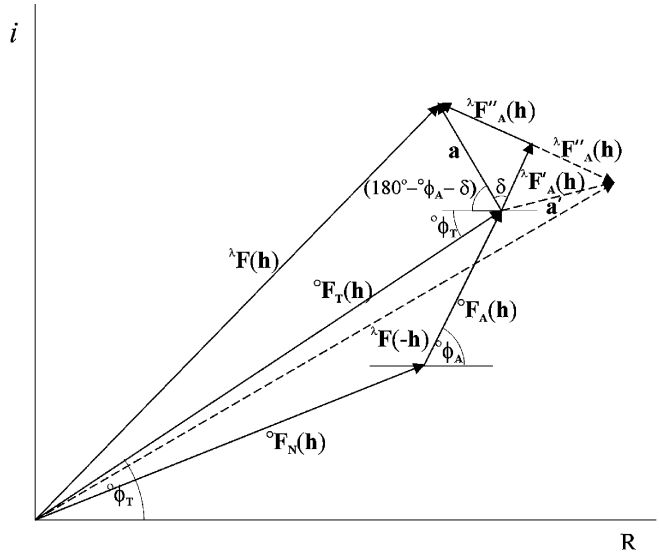


Fig. 5.8 Schematic drawing of structure factors of a biological macromolecule that contains one kind of anomalous scatterer.

Substituting this in Eq. (5.41), we obtain the expression for ${}^{\lambda}F(\mathbf{h})^2$. ${}^{\lambda}F(-\mathbf{h})^2$ is determined from the triangle formed by the vectors ${}^{\lambda}\mathbf{F}(-\mathbf{h})$, ${}^{\circ}\mathbf{F}_T(\mathbf{h})$ and $\mathbf{a}'$ (Fig. 5.8) using a similar approach. Maximum anomalous scattering effects can be expected in intensity differences of reflections that would be equal for exclusively normal scattering. This is the case for Friedel pairs, **h** and –**h**, or their rotational symmetry partners, and the relationship for such differences is given in Eq. (5.27). Of further interest are dispersive differences between structure amplitudes at different wavelengths (Eq. 5.47):

$$\Delta F_{\Delta\lambda} \equiv {}^{\lambda i}F(\mathbf{h}) - {}^{\lambda j}F(\mathbf{h}) \tag{5.47}$$

The anomalous or dispersive intensity differences can be used to determine the structure of the anomalous scatterers. The methods used are the same as for isomorphous replacement, and these will be discussed in the next section.

5.2.7 Determination of the Absolute Configuration

As anomalous scattering destroys the centro-symmetry of the diffraction data, this effect can be used to determine the absolute configuration of chiral biological macromolecules. The most common method is to calculate protein phases based on both hands of the heavy atom or anomalous scatterer structures, and to check the quality of the relevant electron density map, which should be better

for the correct hand. Furthermore, secondary structural elements in proteins (consisting of L-amino acids) such as α-helices should be right-handed.

5.3 Determination of Heavy-Atom Positions

5.3.1 Vector Verification Procedures

A key step in solving the phase problem is to determine the heavy atom or anomalous scatterer positions by analyzing the respective isomorphous, dispersive, or anomalous difference Patterson maps. In simple circumstances, this can be done by an individual inspection of the difference Patterson maps, and this method has been applied often in the past. We will explain this approach with an example. Our crystal has the space group $P2_12_12_1$ and three independent molecules in the asymmetric unit with one heavy atom each (Fig. 5.9). Each atom with coordinates x_j, y_j, z_j has symmetry mates at $-x_j + 1/2, -y_j, z_j + 1/2$; $x_j + 1/2, -y_j + 1/2, -z_j$ and $-x_j, y_j + 1/2, -z_j + 1/2$ according to the crystallographic symmetry. Vectors between symmetry mates adopt the following forms: $\mathbf{x}_i' - \mathbf{x}_i = -2x_i, -2y_i, 0$; $\mathbf{x}_i'' - \mathbf{x}_i = 1/2, -2y_i + 1/2, -2z_i$; $\mathbf{x}_i''' - \mathbf{x}_i = -2x_i, 1/2, -2z_j + 1/2$. One can see that all these vectors lie in special sections of the difference Patterson map, namely in $w = 0$; $u = 1/2$ and $v = 1/2$, respectively. Such vectors are denoted as Harker vectors, and the sections as Harker sections. It can also be seen that the coordinates of the corresponding heavy atom can be determined from the Harker vectors. In Figure 5.9, we have three copies of a molecule with one heavy atom per molecule in the asymmetric unit. These molecules are related by noncrystallographic symmetry (NCS) among each other. One can independently determine the coordinates of each atom, but these coordinates are ambiguous with respect to the different origins that can be chosen for the space group representation. An unambiguous determination is achieved if vectors between heavy atoms within the asymmetric unit are also considered. As such vectors generally do not lie in any special section of the Patterson map, it will be more difficult to locate them. Furthermore, the number of peaks in a heavy-atom Patterson map equals the square of the number of heavy atoms, N, including the origin peak. This means that the interpretation of more complex heavy-atom structures becomes increasingly complicated, and automated methods for solving difference Patterson functions must be applied. These methods may be divided into two categories: vector search methods (e.g., Terwilliger et al., 1987; Steigemann, 1991; Knight, 2000), and superposition methods (Buerger, 1959; Sheldrick, 1991).

We will shortly explain the vector search method. In Figure 5.9, the full lines between heavy atoms represent Harker vectors, the dashed ones cross vectors generated by application of NCS, and the dotted ones vectors generated by combining NCS and space group symmetry (SGS). The complete vector set for the Harker (3×12) and NCS-cross (2×12) vectors is shown in Figure 5.9. Only two

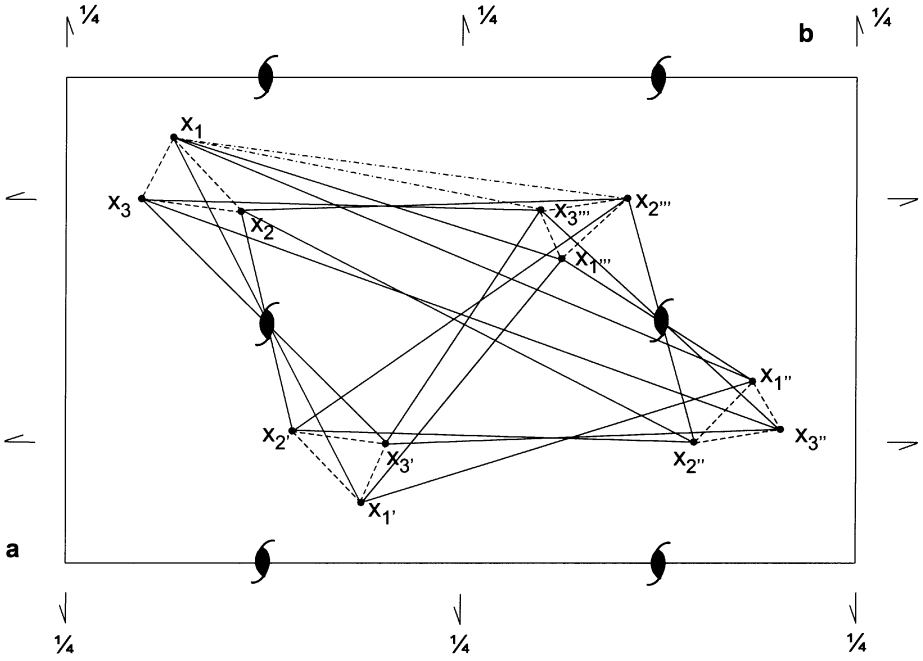


Fig. 5.9 Interatomic vectors between heavy atoms in space group $P2_12_12_1$. There are three copies of a molecule with one heavy atom each in the asymmetric unit. Harker vectors are displayed as full lines, NCS-cross vectors as dashed lines and NCS-SGS-cross vectors (two vectors out of 84 shown only) as dotted lines.

of the 84 NCS-SGS-cross vectors are displayed. Each line stands for the positive and negative interatomic vector. If the NCS operations are known, these can be incorporated into the search routines. As normally the NCS operations (especially the translational components) are unknown at the beginning of the phase determination, they will not be included in the following discussion. The first step is a single site search. An atomic position $\mathbf{x}$ is scanned through the whole unit cell and for each $\mathbf{x}$ vectors $\mathbf{u}_{ij}(\mathbf{x})$ are calculated according to

$$\mathbf{u}_{ij}(\mathbf{x}) = R_i^{\mathrm{SGS}}\mathbf{x} + \mathbf{t}_i^{\mathrm{SGS}} - (R_j^{\mathrm{SGS}}\mathbf{x} + \mathbf{t}_j^{\mathrm{SGS}}) = (R_i^{\mathrm{SGS}} - R_j^{\mathrm{SGS}})\mathbf{x} + \mathbf{t}_i^{\mathrm{SGS}} - \mathbf{t}_j^{\mathrm{SGS}} \,. \tag{5.48}$$

R^{SGS} and $\mathbf{t}^{\mathrm{SGS}}$ are the relevant rotational and translational space group symmetry operators. In our example, i and j extend from 1 to 4. Values for $i = j$ are not considered because they represent self vectors of lengths zero. Thus, 12 Harker vectors are calculated for each position $\mathbf{x}$, their relevant Patterson function value acquired, and a correlation function determined. This correlation function can be the sum, the product, or the minimum of the respective Patterson function values. The sum function is rather insensible against missing

peaks, and is the opposite to the latter two scoring functions. Although the multiplication scoring function is highly selective, in isomorphous, dispersive or anomalous difference Patterson maps some peaks may be absent, and therefore the sum scoring function often gives the best result. As we have three heavy or anomalous scatterer atoms in the asymmetric unit, there should be three symmetry-independent maxima in the vector search. As they occur repeatedly due to several possible origins in the unit cell, it is necessary to find their common origin. This is done in a second step by a cross-vector search. In space group $P1$ the only symmetry operation is the identity, and therefore we do not have Harker vectors. As the choice of the origin in $P1$ is free, one can choose the position of one heavy atom arbitrarily. For a cross-vector search one or several known atomic positions are kept fixed and for each position $\mathbf{x}$ vectors $\mathbf{u}_{ijm}(\mathbf{x})$ and the score functions are calculated. The vectors adopt the form

$$\mathbf{u}_{ijm}(\mathbf{x}) = R_i^{\mathrm{SGS}}\mathbf{x} + \mathbf{t}_i^{\mathrm{SGS}} - (R_j^{\mathrm{SGS}}\mathbf{x}_m^{\mathrm{fix}} + \mathbf{t}_j^{\mathrm{SGS}}) = R_i^{\mathrm{SGS}}\mathbf{x} - R_j^{\mathrm{SGS}}\mathbf{x}_m^{\mathrm{fix}} + \mathbf{t}_i^{\mathrm{SGS}} - \mathbf{t}_j^{\mathrm{SGS}} \,. \tag{5.49}$$

Such a cross-vector search should have peaks for all heavy atoms with respect to a common origin in the unit cell, provided that the quality of the respective difference Patterson is sufficiently good. If the interatomic vectors between two or more heavy atoms are known, one can determine the translational vector $\mathbf{t}^x$ by using it as a scan parameter to calculate the vectors

$$\begin{aligned}\mathbf{u}_{ijmn}(\mathbf{t}^x) &= R_i^{\mathrm{SGS}}(\mathbf{x}_m^{\mathrm{fix}} + \mathbf{t}^x) + \mathbf{t}_i^{\mathrm{SGS}} - (R_j^{\mathrm{SGS}}(\mathbf{x}_n^{\mathrm{fix}} + \mathbf{t}^x)) + \mathbf{t}_j^{\mathrm{SGS}}) \\ &= (R_i^{\mathrm{SGS}} - R_j^{\mathrm{SGS}})\mathbf{t}^x + (R_i^{\mathrm{SGS}}\mathbf{x}_m - R_j^{\mathrm{SGS}}\mathbf{x}_n) + (\mathbf{t}_i^{\mathrm{SGS}} - \mathbf{t}_j^{\mathrm{SGS}})\end{aligned} \tag{5.50}$$

and the relevant scoring function values. Vector search procedures based on these principles and including possible knowledge of NCS are provided in the program systems PROTEIN (Steigemann, 1991), SOLVE (Terwilliger and Berendzen, 1999), and the CCP4 program RSPS (Knight, 2000).

5.3.2 Direct Methods

The term "direct methods" denotes those methods which try to derive the structure factor phases directly from the observed amplitudes through mathematical relationships. Direct methods, implemented in widely used highly automated computer programs such as MULTAN (Main et al., 1980) and SHELXS (Sheldrick, 1990) provide computationally efficient solutions for structures with fewer than about 100 independent non-H atoms. Pushing up the limits to about 2000 independent non-H atoms using exclusively native data was achieved by the development of a direct methods procedure (Weeks et al., 1993) that has come to be known as *Shake-and-Bake*. The principles and applications of direct methods in X-ray crystallography of biological macromolecules has been reviewed (Sheldrick et al., 2001) and served as the basis for this section.

For direct methods it is necessary to replace the usual structure factors $\mathbf{F_h}$ by the normalized structure factors (Hauptman and Karle, 1953),

$$\mathbf{E_h} = |\mathbf{E_h}| \exp i\varphi_{\mathbf{h}}$$
$$|\mathbf{E_h}| = \frac{|\mathbf{F_h}|}{\langle |\mathbf{F_h}|^2 \rangle^{1/2}} = \frac{C\langle \exp[-B(\sin\theta)^2/\lambda^2] \rangle^{-1} |\mathbf{F_h}|_{\text{obs}}}{\left(\varepsilon_{\mathbf{h}} \sum_{j=1}^{N} f_j^2 \right)^{1/2}} , \tag{5.51}$$

where the angle brackets indicate probabilistic or statistical expectation values, $|\mathbf{E_h}|$ and $\varphi_{\mathbf{h}}$ are the amplitude and phase angle of the normalized structure factor, C and B are the scale and temperature factors from the Wilson plot, and the $\varepsilon_{\mathbf{h}} \geq 1$ are factors that account for multiple enhancement for certain special reflection classes due to space group symmetry (Shmueli and Wilson, 2001). The other symbols have the known meanings from the familiar structure factor equation. The subscript for the indices vector $\mathbf{h}$ has been chosen for clarity in the subsequent mathematical relationships. The condition $\langle |\mathbf{E}|^2 \rangle = \mathbf{1}$ is always imposed, and the values for $\langle |\mathbf{E_h}| \rangle$ are constant for all concentric resolution shells unlike $\langle |\mathbf{F_h}| \rangle$, which decreases with increasing $\sin\theta/\lambda$. Thus, the normalization process places all reflections on a common basis, and this is a great advantage with regard to the probability distributions that form the basis of direct methods.

For determining the positions of heavy atoms or anomalous scatterers, the respective isomorphous, dispersive, or anomalous difference amplitudes are used in direct methods. However, they must be in the form of normalized difference structure factor magnitudes $|\mathbf{E}_\Delta|$. This can be done with the programs from Blessing's data reduction and error analysis routines (DREAR): LEVY and EVAL for structure factor normalization according to Eq. (5.51) (Blessing et al., 1996), LOCSCL for local scaling of the SIR and SAS magnitudes (Matthews and Czerwinski, 1975; Blessing, 1997) and DIFFE for the determination of the actual difference magnitudes (Blessing and Smith, 1999).

SIR and SAS differences are calculated as greatest lower bound estimates according to Eqs. (5.52) and (5.53):

$$|\mathbf{E}_\Delta|_{\text{SIR}} = \frac{\left| \left(\sum_{j=1}^{N_{\text{der}}} |f_j|^2 \right)^{1/2} |\mathbf{E}_{\text{der}}| - \left(\sum_{j=1}^{N_{\text{nat}}} |f_j|^2 \right)^{1/2} |\mathbf{E}_{\text{nat}}| \right|}{q \left[\left(\sum_{j=1}^{N_{\text{der}}} |f_j|^2 \right) - \left(\sum_{j=1}^{N_{\text{nat}}} |f_j^2| \right) \right]^{1/2}} , \tag{5.52}$$

where $|\mathbf{E}_{\text{nat}}|$ and $|\mathbf{E}_{\text{der}}|$ are the normalized structure factor magnitudes of the native and derivative data set, respectively, $|f_j| = |f_j^0 + f_j' + if_j''| = [(f_j^0 + f_j')^2 + (f_j'')^2]^{1/2}$ are the atomic scattering factors which allow for the possibility of anomalous scattering, $q = q_0 \exp(q_1 S^2 + q_2 S^4)$ is a least-squares fitted empirical

scaling function dependent on $S = \sin\theta/\lambda$ that imposes the condition $\langle|\mathbf{E}_\Delta|^2\rangle = 1$ and is used to define q_0, q_1 and q_2.

$$|\mathbf{E}_\Delta|_{\mathrm{SAS}} = \frac{\left[\sum_{j=1}^{N}(f_j^0 + f_j')^2 + (f_j'')^2\right]^{1/2} \; ||\mathbf{E}_{+\mathbf{h}}| - |\mathbf{E}_{-\mathbf{h}}||}{2q\left[\sum_{j=1}^{N}(f_j'')^2\right]^{1/2}} \; . \tag{5.53}$$

$|\mathbf{E}_{+\mathbf{h}}|$ and $|\mathbf{E}_{-\mathbf{h}}|$ are the normalized structure factor magnitudes of the Friedel pairs, and q is again an empirical renormalization scaling function that imposes the condition $\langle|\mathbf{E}_\Delta|^2\rangle = 1$.

In general, the phase and the amplitude of a wave are independent quantities. However, in X-ray diffraction there exist relationships between them, and these result from two important properties of atomic structures. Their electron density is everywhere positive – that is, $\rho(\mathbf{r}) \geq 0$ (positivity) – and they are composed of discrete atoms (atomicity). Based on these properties, magnitude-dependent entities – which are linear combinations of phases called structure invariants – have been derived. They were named in this way because these quantities are independent of the choice of the origin. The most useful of the structure invariants are the three-phase or triplet invariants

$$\Phi_{\mathbf{hk}} = \varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{-\mathbf{h}-\mathbf{k}} \; . \tag{5.54}$$

Its conditional probability distribution (Cochran, 1955) is

$$P(\Phi_{\mathbf{hk}}) = [2\pi I_0(A_{\mathbf{hk}})]^{-1} \exp(A_{\mathbf{hk}} \cos \Phi_{\mathbf{hk}}) \; , \tag{5.55}$$

where

$$A_{\mathbf{hk}} = (2/N^{1/2})|\mathbf{E}_{\mathbf{h}}\mathbf{E}_{\mathbf{k}}\mathbf{E}_{\mathbf{h}+\mathbf{k}}| \tag{5.56}$$

N is the number of atoms, here presumed to be identical, in the asymmetric unit of the corresponding primitive unit cell, and I_0 is a modified Bessel function. Estimates of the invariant values are most reliable when the normalized structure factor magnitudes $|\mathbf{E}_{\mathbf{h}}|, |\mathbf{E}_{\mathbf{k}}|$ and $|\mathbf{E}_{-\mathbf{h}-\mathbf{k}}|$ are large and the number of atoms N in the unit cell is small. This is the main reason why direct phasing is more difficult for macromolecules than it is for small molecules.

The phasing procedure starts in sorting the $|\mathbf{E}|$ values with respect to their magnitude, and a list of invariants with their joint probabilities $A_{\mathbf{hk}}$ is calculated using the largest $|\mathbf{E}|$ values. The invariants with the largest $A_{\mathbf{hk}}$s are retained. Now, a start-phasing set is still needed to initiate the calculation of phases by means of the invariants. Depending on the space group, a small number of phases can be assigned arbitrarily in order to fix the origin position and, in non-centro-symmetric space groups, the enantiomer. However, these reflections

provide an inadequate basis for subsequent phase development. In order to extend the starting phase set, reflections are assigned a couple of different phase values. Phases are developed then by running all possible combinations of these phases in a multisolution approach (Germain and Woolfson, 1968), in the hope that the correct solution is included. Solutions must be identified on the basis of some suitable figure of merit.

Once a set of phases has been chosen, it must be refined against the set of structure invariants whose values are presumed known. This step is denoted as shaking, and phase refinement or expansion take place in reciprocal space. So far, only two optimization methods (tangent refinement and parameter shift optimization of the minimal function) have been proven to be of practical value. The tangent formula

$$\tan(\varphi_{\mathbf{h}}) = \frac{-\sum_{\mathbf{k}} |\mathbf{E}_{\mathbf{k}}\mathbf{E}_{-\mathbf{h}-\mathbf{k}}| \sin(\varphi_{\mathbf{k}} + \varphi_{-\mathbf{h}-\mathbf{k}})}{\sum_{\mathbf{k}} |\mathbf{E}_{\mathbf{k}}\mathbf{E}_{-\mathbf{h}-\mathbf{k}}| \cos(\varphi_{\mathbf{k}} + \varphi_{-\mathbf{h}-\mathbf{k}})} \tag{5.57}$$

(Karle and Hauptman, 1956) is used in conventional direct-methods programs, and also in the phase refinement part of the dual-space Shake-and-Bake procedure (Weeks et al., 1994; Sheldrick and Gould, 1995) to calculate $\varphi_{\mathbf{h}}$, given a sufficient number of pairs $(\varphi_{\mathbf{k}}, \varphi_{-\mathbf{h}-\mathbf{k}})$ of known phases. The estimate of $\varphi_{\mathbf{h}}$ by the tangent formula is only reliable for $|\mathbf{E}_{\mathbf{h}}| >> 1$ and for structures with a limited number of atoms N.

Another possibility for phase refinement or phase expansion is the constrained minimization of an objective function such as the minimal function (Debaerdemaker and Woolfson, 1983; Hauptman, 1991):

$$R(\Phi) = \sum_{\mathbf{h},\mathbf{k}} A_{\mathbf{hk}} \{\cos \Phi_{\mathbf{hk}} - [I_1(A_{\mathbf{hk}})/I_0(A_{\mathbf{hk}})]\}^2 / \sum_{\mathbf{h},\mathbf{k}} A_{\mathbf{hk}} \ . \tag{5.58}$$

$R(\Phi)$ is a measure of the mean square difference between the values of the triplets calculated using a particular set of phases, and the expected values of the same triplets as given by the ration of the modified Bessel functions. The minimal function is expected to have a constrained global minimum when the phases are equal to their correct values for some choice of origin and enantiomer. It transpired that the minimal function is also an extremely useful figure of merit.

In principle, any minimization technique could be used to minimize $R(\Phi)$ by varying the phases. So far, a seemingly simple algorithm, known as parameter shift (Bhuiya and Stanley, 1963), has proven to be quite powerful and efficient when used within the Shake-and-Bake context to reduce the minimal function, which is not described in detail here.

The exploitation of real space constraints due to the atomicity for phasing is called baking. Automatic real-space electron density map interpretation in the Shake-and-Bake procedure consists of selecting an appropriate number of the largest peaks in each cycle to be used as an updated trial structure, without re-

gard to chemical constraints other than a minimum allowed distance between atoms. Selecting the largest N_u peaks (N_u = number of atoms, heavy atoms or anomalous scatterers per asymmetric unit) for a true small molecule or for heavy-atom or anomalous scatterer substructures delivers satisfying results. The trial structure is transformed back to reciprocal space and subjected to phase refinement by the tangent formula.

The Shake-and-Bake algorithm has been implemented independently in the two computer programs SnB (Miller et al., 1994; Weeks and Miller, 1999) and SHELXD (Sheldrick, 1997, 1998). The determination of the anomalous scatterers from MAD data with program SnB will be illustrated in Part II of this book.

5.4 Phase Calculation

5.4.1 Refinement of Heavy-Atom Parameters

Before the protein phases can be calculated, it is necessary to refine the heavy-atom parameters. These are the coordinates x, y, z, the temperature factor (either isotropic or anisotropic), and the occupancy. The refinement modifies the parameters in such a way that $|\mathbf{F}_{PH}(obs)|$ becomes as close as possible to $|\mathbf{F}_{PH}(calc)|$. Using the method of least squares, the refinement according to Rossmann (1960) minimizes Eq. (5.59):

$$\varepsilon = \sum_{\mathbf{h}} w(\mathbf{h})[(F_{PH} - F_P)^2 - kF_{H\,calc}^2]^2 \tag{5.59}$$

where k is a scaling factor to correct $F_{H\,calc}^2$ to a theoretically more acceptable value because, according to Eq. (5.4), $\mathbf{F}_{PH}-\mathbf{F}_P$ and $\mathbf{F}_H$ have approximately the same length when $\mathbf{F}_{PH}$, $\mathbf{F}_P$ and $\mathbf{F}_H$ point in the same direction. The probability for this case will be high if the difference between F_{PH} and F_P is large. An improvement can be obtained if the contribution from the anomalous scattering is included (Dodson and Vijayan, 1971).

For the parameter refinement of anomalous scattering sites, the differences between the observed and calculated structure factor amplitudes for ${}^{\circ}F_A$ are subjected to minimization. Another approach treats the anomalous or dispersive contributions as in MIR phasing.

From the refined heavy-atom parameters, preliminary protein phase angles α_P can be obtained, as shown in the corresponding Harker construction. A further refinement of the heavy-atom parameters can be achieved by the "lack of closure" method (Dickerson et al., 1968) incorporating this knowledge. The definition of "lack of closure" ε is illustrated in Figure 5.10a and b. In the case of perfect isomorphism, the vector triangle $\mathbf{F}_P+\mathbf{F}_H=\mathbf{F}_{PH}$ closes exactly (Fig. 5.10a). In

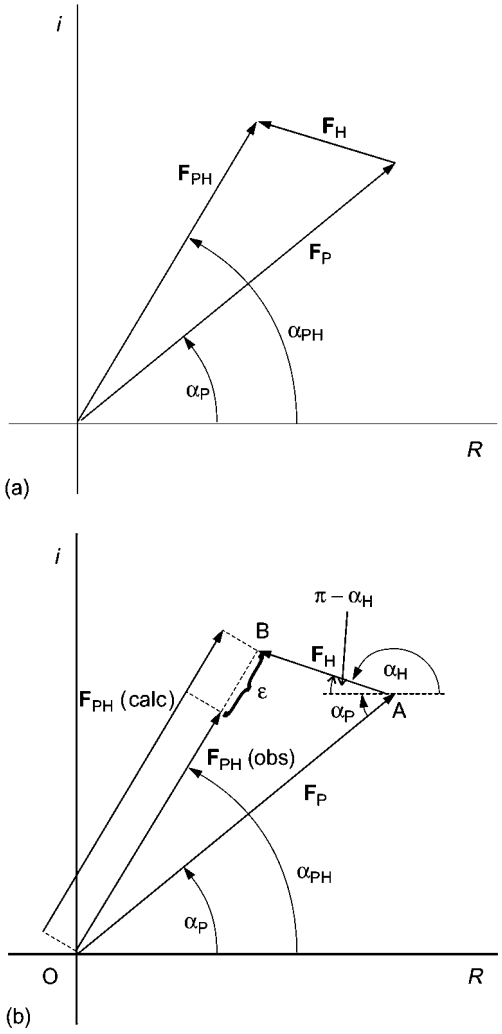


Fig. 5.10 Definition of "lack of closure". (a) Perfect isomorphism; (b) usually, the observed and calculated values for $\mathbf{F}_{PH}$ differ by the "lack of closure" ε.

practice, this condition will not be fulfilled and a difference ε between the observed F_{PH} and the calculated F_{PH} will remain (Fig. 5.10b). F_{PH}(calc) can be obtained from the triangle OAB (Fig. 5.10b) with the cosine rule:

$$F_{PH} = [F_P^2 + F_H^2 + 2F_P \times F_H \cos(\alpha_H - \alpha_P)]^{1/2} \tag{5.60}$$

The function which is minimized by the least-squares method is:

$$E_j = \sum_{\mathbf{h}} m_{\mathbf{h}} \varepsilon_j(\mathbf{h})^2 \tag{5.61}$$

where

$$\varepsilon_j = k_j F_{\mathrm{PH}j}(\mathrm{obs}) - F_{\mathrm{PH}j}(\mathrm{calc}) \tag{5.62}$$

is the "lack of closure" for the heavy-atom derivative j, k_j is a scaling factor, and $m_{\mathbf{h}}$ is a weighting factor.

5.4.2 Protein Phases

As the structure factor amplitudes F_{P}, F_{PH}, F_{H} and a_{H} are known, the protein phase angle a_{P} can be calculated. For the single isomorphous replacement situation (see Fig. 5.1), ε is zero only for the two protein phase angles a_{P} where the two circles for F_{P} and F_{PH} intersect. In practice, all these observed quantities exhibit errors. For the treatment of these errors it is assumed that all errors are in $\mathbf{F}_{\mathrm{PH}}$ and that both $\mathbf{F}_{\mathrm{H}}$ and $\mathbf{F}_{\mathrm{P}}$ are error-free. For each protein phase angle a, $\varepsilon(a)$ is calculated. The smaller $\varepsilon(a)$ is, the higher is the probability of a correct phase angle a. For each reflection of a derivative j a Gaussian probability distribution is assumed for ε according to Eq. (5.63):

$$P(a) = P(\varepsilon) = N \exp\left[-\frac{\varepsilon^2(a)}{2E^2}\right] \tag{5.63}$$

where N is a normalization factor and E^2 the mean square value of ε. Small values of E are related to probability curves with sharp peaks and well-determined phase angles and the opposite is true for large E values. Such phase-angle probability curves can be calculated for each individual reflection and derivative. For single isomorphous replacement this curve is symmetric, with two high peaks corresponding to the two possible solutions for a_{P}. We obtain the total probability for each reflection with contributions from n heavy-atom derivatives by multiplying the individual probabilities:

$$P(a) = \prod_{j=1}^{n} P_j(a) = N' \exp\left[-\sum_j \frac{\varepsilon_j^2(a)}{2E_j^2}\right] \tag{5.64}$$

These curves will be nonsymmetric with one or several maxima (see Fig. 5.11a and b).

The question arises of which phases should be taken in the electron density equation to calculate the best electron density function. An immediate guess would be to use the phases where $P(a)$ has the highest value. This approach would be appropriate for unimodal distributions, but not for bimodal distributions. Blow and Crick (1959) derived the phase value that must be applied under the assumption that the mean square error in electron density over the unit cell is minimal. For one reflection this is given by:

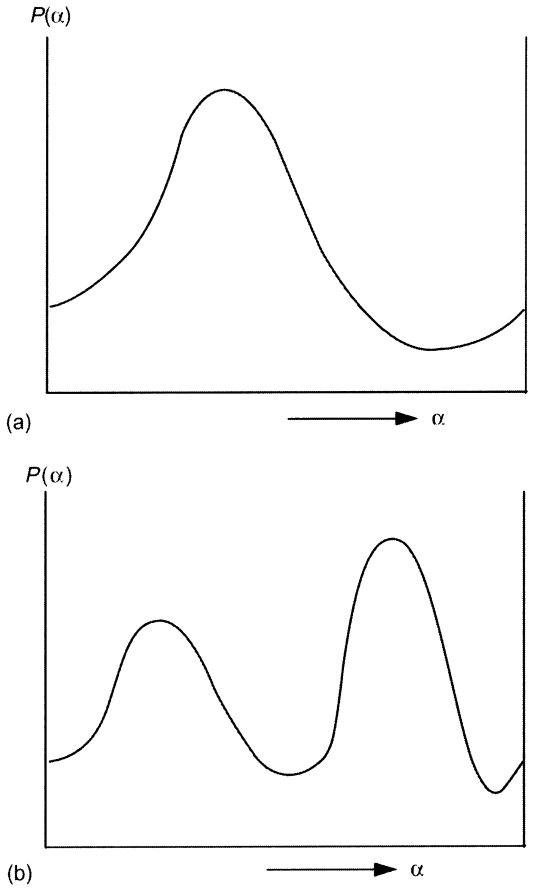


Fig. 5.11 Total probability curves $P(\alpha)$ for two different reflections. (a) For one derivative; (b) for more than one derivative.

$$\langle \Delta\rho^2 \rangle = \frac{1}{V^2} (\mathbf{F}_s - \mathbf{F}_t)^2 \tag{5.65}$$

where $\mathbf{F}_t$ is the true and $\mathbf{F}_s$ the structure factor applied in the Fourier synthesis. The mean square error is then obtained as:

$$\langle \Delta\rho^2 \rangle = \frac{1}{V^2} \frac{\int_{\alpha=0}^{2\pi} (\mathbf{F}_s - F \exp i\alpha)^2 P(\alpha) \mathrm{d}\alpha}{\int_{\alpha=0}^{2\pi} P(\alpha) \mathrm{d}\alpha} \tag{5.66}$$

$\mathbf{F}_t$ has a phase probability of $P(\alpha)$, and has been given as $\mathbf{F}_t = F \exp i\alpha$. It can be shown that the numerator integral in Eq. (5.66) is minimal if:

$$\mathbf{F}_{s(best)} = F\frac{\int_{\alpha=0}^{2\pi} \exp(i\alpha)P(\alpha)\mathrm{d}(\alpha)}{\int_{\alpha=0}^{2\pi} P(\alpha)\mathrm{d}\alpha} mF\exp(i\alpha_{best}) \tag{5.67}$$

Equation (5.67) corresponds to the center of gravity of the probability distribution with polar coordinates (mF, α_{best}), where m is defined as magnitude of $\mathbf{m}$ given by:

$$\mathbf{m} = \frac{\int_{\alpha=0}^{2\pi} P(\alpha)\exp(i\alpha)\mathrm{d}\alpha}{\int_{\alpha=0}^{2\pi} P(\alpha)\mathrm{d}\alpha} \tag{5.68}$$

This magnitude of $\mathbf{m}$ is equivalent to a weighting function and is designated the "figure of merit". The electron density map calculated with mF and α_{best} is known as "best Fourier", and should represent a Fourier map with minimum least-squares error from the true Fourier map.

For the total error of the "best Fourier", the following equation has been derived:

$$\langle \Delta\rho^2 \rangle = \frac{1}{V^2}\sum_{\mathbf{h}} F^2(\mathbf{h})(1-m^2) \tag{5.69}$$

The order of magnitude of this error may be illustrated by the example of the structure determination of lysozyme. The root mean square error in the Fourier synthesis was 0.35 $Å^{-3}$ with values of 2.0 Å resolution for the diffraction data and a mean "figure of merit" of 0.6.

The program systems CCP4 (CCP4, 1994) and PROTEIN (Steigemann, 1991) contain all routines necessary to calculate protein phases according to the MIRAS technique and a number of different kinds of Fourier maps. Alternative probabilistic approaches for the phase calculation are used in programs MLPHARE (Otwinowski, 1991) and SHARP (de La Fortelle and Bricogne, 1997). Both programs can also carry out MAD phasing. The MADSYS program (Hendrickson, 1991) is based on the algebraic approach outlined in Section 5.2.6, and executes all tasks of a MAD analysis, from scaling to phase-angle calculation.

5.4.3 Maximum-Likelihood Parameter Refinement and Phase Calculation

As the program SHARP (de La Fortelle and Bricogne, 1997) is now a standard approach to heavy atom and anomalous scatterer parameter refinement and phase calculation for MIR and MAD methods, the basic principles of the underlying maximum-likelihood (ML) formalism is briefly outlined here. The so-far used least-squares (LS) model is always formulated as a prescription for turning given values of model parameters into "calculated" (error-free) values to be compared with the observables. Error estimates are obtained *a posteriori* by examining the residual discrepancy between the "calculated" and "observed" quantities. Thereby, a bias is introduced in the model incorporating some degree of randomness whenever a distribution for a random quantity is replaced by a value for that quantity. On the other hand, a likelihood-based model puts its predictions directly in the form of probability distributions for the observables, the quantities called "calculated" in the LS formalism usually appearing as parameters in these distributions.

In the LS formalism, the calculated structure factor amplitude $|\mathbf{F}_{\mathrm{PH}j}(\mathbf{h})|$, which will be denoted as r_j, is given (allowing a scale factor k_j) by (Eq. 5.70):

$$r_j = k_j|\mathbf{F}_{\mathrm{P}}(\mathbf{h}) + \mathbf{F}_{\mathrm{H}j}(\mathbf{h})| \; . \tag{5.70}$$

Nonisomorphism is estimated *a posteriori* by an analysis of lack-of-closure errors, but the actual structure of the LS equations precludes its refinement. Furthermore, no attention is paid to the interaction between the nonisomorphism variance and the relative scale of native and derivative structure factors.

In the ML formalism, the introduction of randomness caused by nonisomorphism is reflected by considering not the values but the distributions of all quantities affected by it. The key step is now to introduce for each jth data set (heavy-atom derivative or MAD data) a so-called perturbed $\mathbf{F}_{\mathrm{P}j}(\mathbf{h}) = \mathbf{F}_{\mathrm{PH}j}(\mathbf{h}) - \mathbf{F}_{\mathrm{H}j}(\mathbf{h})$. A native data set, if present, is treated as compound $j = 1$ with zero heavy-atom contribution and zero nonisomorphism. Under lattice-preserving nonisomorphism $\mathbf{F}_{\mathrm{P}j}(\mathbf{h})$ is a random complex number with mathematical expectation

$$\langle \mathbf{F}_{\mathrm{P}j}(\mathbf{h}) \rangle = \mathbf{F}_{\mathrm{P}*}(\mathbf{h}) \times D_j \tag{5.71}$$

and variance V_j^{r}. Here, the complex number $\mathbf{F}_{\mathrm{P}*}(\mathbf{h})$ denotes the unperturbed native structure factor, which is not directly observable without error, but whose (hidden) value acts as a common parameter linking the distributions of all $\mathbf{F}_{\mathrm{P}j}(\mathbf{h})$ for the various js. The quantity D_j is an "attenuation factor", the value of which lies between 0 and 1, and is connected to the variance parameter V_j^{r} in such a way as to keep the expected value in resolution shells independent of the degree of nonisomorphism. Substitution of Eq. (5.71) into Eq. (5.70) shows that $\mathbf{F}_{\mathrm{P}j}(\mathbf{h})$ is now a random complex number, the distribution of which is a two-dimensional Gaussian of variance V_j^{r} centered at $\langle \mathbf{F}_{\mathrm{P}j}(\mathbf{h}) \rangle$.

If, for each reflection **h**, we could measure $\mathbf{F}_{Pj}(\mathbf{h})$ as a complex number, then the probability of measuring a complex value $R_j \exp i\psi_j$ (for prescribed values of the parameters $r_j \exp i\varphi_j$ and V_j^r) would be:

$$p(R_j \exp i\psi_j | r_j \exp i\varphi_j, V_j^r) = \frac{1}{2\pi V_j^r} \exp\left[-\frac{|r_j \exp i\varphi_j - R_j \exp i\psi_j|^2}{2V_j^r} \right] \tag{5.72}$$

Since one can measure only structure factor amplitudes, there is a need to derive from the previous formula the probability of measuring the amplitude R_j, whatever its phase may be.

This is done by analytical integration of the previous distribution over the phase, to obtain Eq. (5.73):

$$p(R_j | \mathbf{F}_{P*}, \mathbf{F}_{Hj}, k_j, V_j^r) = \mathscr{R}(r_j(\mathbf{F}_{P*}, \mathbf{F}_{Hj}, k_j, V_j^r), R_j, V_j^r) \ , \tag{5.73}$$

where $\mathscr{R}$ (Eq. 5.74) denotes the Rice distribution (Rice, 1944):

$$\mathscr{R}(r, R, V) = \frac{R}{V} \exp\left(-\frac{r^2 + R^2}{2V} \right) I_0\left(\frac{rR}{V} \right) \ . \tag{5.74}$$

If there are M distinct isomorphous derivatives with statistically uncorrelated lack-of-isomorphism errors, the conditional joint probability distribution of the M structure-factor amplitudes for these compounds is:

$$p(R_1, \ldots, R_M | r_1, V_1^r, \ldots, r_M, V_M^r) = \prod_{j=1}^{M} \mathscr{R}(r_j(\mathbf{F}_{Pj}, \mathbf{F}_{Hj}, k_j, V_j^r), R_j, V_j^r) \ . \tag{5.75}$$

One could now go over to the likelihood by substituting the measured values R_j^{obs} for the relevant measurable parameters R_j. However, the physical measurement leads in this case only to a Gaussian probability distribution $\mathscr{P}^{obs}(R_j^{obs} | \hat{R}_j, \sigma_j^R)$, characterized by its mean $\hat{R}_j$ and its standard deviation σ_j^R. The likelihood of the current heavy-atom model is obtained by the classical, value-based, likelihood function over all possible values of the measurement R_j^{obs}, each weighted by its observational probability.

We will not further explain the ML method in detail (this is done by de La Fortelle and Bricogne, 1997, the first part having been used for this paragraph). Likelihoods are developed on this basis for the heavy-atom model of isomorphous replacement and the anomalous scatterer model of the MAD measurements, and these are combined to a common likelihood function which is made maximal by refining the relevant parameters. Thereby, a suitable parameterization for the derivative structure factors, scale factors and lack-of-isomorphism variances had to be performed. The refined global parameters are then used to calculate best Fourier phases and Hendrickson–Lattman coefficients, which can be used for phase combination purposes (for an explanation, see Section 6.7).

The Fourier phases can be used to calculate residual maps to detect further heavy atoms or anomalous scatterers.

5.4.4 Cross-Phasing of Heavy-Atom Derivatives or Anomalous Dispersion Data

It is often the case that several heavy-atom derivatives have been prepared and respective X-ray data sets have been collected, but only for one derivative can the heavy-atom positions be determined reliably. The heavy-atom positions of the other derivatives, provided that they are isomorphous, can then be determined from a difference Fourier map of the following form:

$$\Delta\rho(\mathbf{r}) = \frac{1}{V}\sum_{\mathbf{h}}(F_{\mathrm{PH}}(\mathbf{h}) - F_{\mathrm{H}}(\mathbf{h}))\exp[-2\pi i\mathbf{hr} + ia_{\mathrm{P}}(\mathbf{h})]\ . \tag{5.76}$$

This is a Fourier summation with the isomorphous differences as coefficients and the phases of the protein. As we have phase information from one derivative only, only the phases of its heavy-atom set a_{H} are available. Using the approximation relationship for the isomorphous difference (Eq. (5.4), it can be shown (see Drenth, 1994, p. 151) that:

$$\Delta\rho(\mathbf{r}) = \frac{1}{2}\,\frac{1}{V}\sum_{\mathbf{h}}F_{\mathrm{H}}(\mathbf{h})\exp[-2\pi i\mathbf{hr} + ia_{\mathrm{H}}(\mathbf{h})]\ . \tag{5.77}$$

Thus, using the isomorphous differences of the other derivative and the SIR phases as protein phases, the corresponding difference Fourier map shows positive electron density at the site of attached atoms, and negative density at the position of removed atoms with half heights. If we take the anomalous or dispersive differences we can determine the positions of anomalous scatterers.

A different type of Fourier synthesis is the already-mentioned residual map. These can be calculated when the structure determination is almost finished, and they adopt the form:

$$\text{Residual Fourier} = \frac{1}{V}\sum_{\mathbf{h}}(F_{\mathrm{PH}}(\mathbf{h}) - |\mathbf{F}_{\mathrm{P}}(\mathbf{h}) + \mathbf{F}_{\mathrm{H}}(\mathbf{h})|)\exp[-2\pi i\mathbf{hr} + ia_{\mathrm{PH}}(\mathbf{h})]\ . \tag{5.78}$$

The phase angles a_{PH} are calculated for the present model of the derivative. Similar Fourier maps can be calculated for anomalous data and used to detect additional minor heavy-atom sites or anomalous scatterers, as mentioned previously.

5.5
Patterson Search Methods (Molecular Replacement)

If the structures of molecules are similar (virtually identical), or they contain a major similar part, this can be used to determine the crystal structure of the related molecule if the structure of the other molecule is known. This is done by systematically exploring the Patterson function of the crystal structure to be determined with the Patterson function of the search model. Let us first consider some important features of the Patterson function. The relationship between two identical molecules in the search crystal structure (Fig. 5.12a) can generally be formulated as:

$$\mathbf{X}_2 = [C]\mathbf{X}_1 + \mathbf{d} \tag{5.79}$$

Equivalent positions $\mathbf{X}_1$ in molecule 1 are at positions $\mathbf{X}_2$ of molecule 2. $[C]$ is the rotation matrix and $\mathbf{d}$ the translation vector of the movement of the molecule. Figure 5.12b shows the Patterson function belonging to the molecular arrangement in Figure 5.12a. It is evident that around the origin vectors are assembled that are intramolecular, whereas the vectors around lines AB and EF are intermolecular. The intramolecular vectors depend on the molecule orienta-

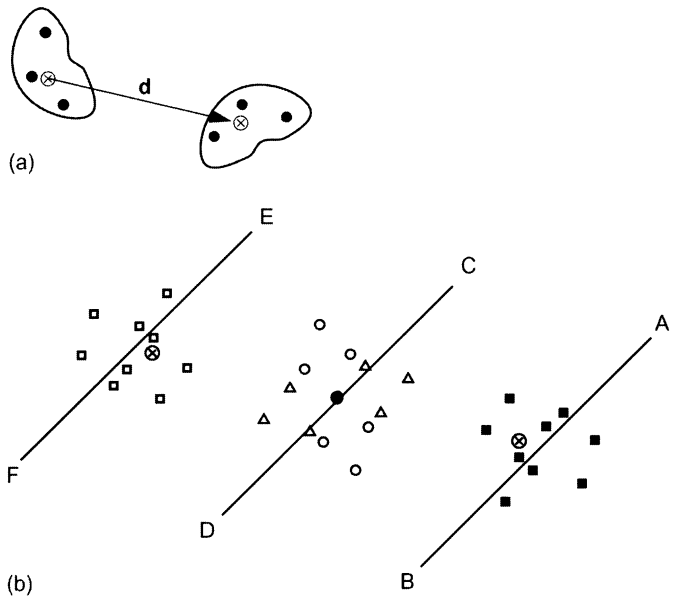


Fig. 5.12 Patterson function of two identical molecules separated by the spatial movement given in Eq. (5.79). (a) Positions of the two molecules; (b) interatomic vectors of structure in (a). △, ○, Intramolecular vectors of the left and right molecule; □, ■, intermolecular vectors.

tion only, and therefore can be used for its determination. Once the orientation of the molecule(s) has been elucidated this can be used to reveal the translation of the molecule(s) by analyzing the intermolecular vector part of the Patterson function. The distinction between intra- and intermolecular vector sets and exploiting them for orientation and translation determination was first realized by Hoppe (1957), while the extension to protein crystallography and the first mathematical formulation of rotation function was provided by Rossman and Blow (1962).

5.5.1 Rotation Function

The intramolecular vectors are arranged in a volume around the origin of the Patterson function with a radius equal to the dimension of the molecule. The rotational search is then carried out in this volume U. The rotation $[C]$ of the molecule is accompanied by a rotation of its corresponding Patterson function $P(\mathbf{u})$ to the rotated position $P_s(\mathbf{u})$. The search Patterson (deduced either from the search model (cross-rotation) or from the crystal Patterson itself (self-rotation)) is rotated to any possible rotational orientation $[C]^{-1}\mathbf{u}$ characterized by three rotational angles, α, β, and γ. When a structural model is available, the target data are calculated by placing the model within a $P1$ unit cell whose dimensions guarantee that U contains only self vectors. The rotational angles may be defined in different ways. The situation for Eulerian angles and spherical polar angles is shown in Figure 5.13a and b, respectively. The now generally used convention for Eulerian angles is illustrated in Figure 5.13a. In the orthogonal coordinate system **x**, **y**, **z**, the first rotation is by the angle α around the **z**-axis, then a rotation by the angle β around the new **y**-axis, and finally a rotation by the angle γ around the new **z**-axis. Spherical polar coordinates (Fig. 5.13b) define a rotation axis κ by the two angles φ and ψ. At each angular position the actual functional values are correlated with those of the crystal Patterson all through the volume u and integrated over this volume. The correlation function may be the sum or the product of each corresponding pair of values. Rossman and Blow (1962) proposed a product function and the rotation function for this case, as given by Eq. (5.80):

$$R(\alpha,\beta,\gamma) = \frac{1}{V}\int_U P_t(\mathbf{u})P_s([C]^{-1}\mathbf{u})\mathrm{d}^3\mathbf{u} \tag{5.80}$$

The function has maxima if the intramolecular vector sets are coincident. The calculation can be carried out in both direct and reciprocal space. In the direct-space formulation (Huber, 1965; Nordman, 1966), an interpolation is needed after each rotation since the values of the Patterson functions are only available at discrete sampling points.

The reciprocal-space formulation of the above integral is obtained by substituting the Patterson functions by their Fourier summations:

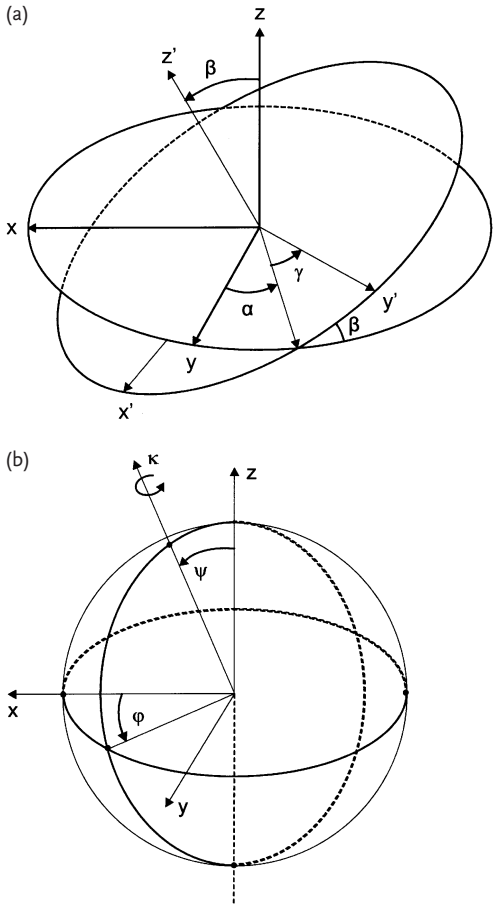


Fig. 5.13 Illustration of rotations defined by: (a) Eulerian angles (α, β, γ); and (b) spherical polar angles (φ, ψ, κ).

$$P(\mathbf{u}) = \frac{1}{V}\sum_{\mathbf{h}} I(\mathbf{h}) \exp(-2\pi i \mathbf{hu}) \; . \tag{5.81}$$

Taking into account that $I(-\mathbf{h}) = I(\mathbf{h})$, one obtains Eq. (5.82):

$$\begin{aligned} R(\alpha,\beta,\gamma) &= \frac{1}{V_t V_s}\sum_{\mathbf{h}}\sum_{\mathbf{k}} I_t(\mathbf{h}) I_s(\mathbf{k}) \frac{1}{V}\int_U \exp[2\pi i(\mathbf{h} - \mathbf{k}[C]^{-1})\mathbf{u}]\mathrm{d}^3\mathbf{u} \\ &= \frac{1}{V_t V_s}\sum_{\mathbf{h}}\sum_{\mathbf{k}} I_t(\mathbf{h}) I_s(\mathbf{k}) \chi_U(\mathbf{h} - \mathbf{k}[C]^{-1}) \; , \end{aligned} \tag{5.82}$$

where χ_U, the interference function, is the Fourier transform of the characteristic function of U – that is, a function that adopts the value 1 within U and 0

outside. In principle, the region of integration could have any shape, but a useful choice for the region is a sphere with radius b. Making the substitution $\mathbf{h} - \mathbf{k}[C]^{-1} = \mathbf{s}$, we obtain

$$\chi_U = (3/4\pi b^3) \int_0^b \int_0^\pi \int_0^{2\pi} \exp(2\pi i \mathbf{su}) u^2 \sin(\theta) \mathrm{d}u \mathrm{d}\theta \mathrm{d}\varphi$$
$$= 3[\sin(2\pi sb) - 2\pi sb \cos(2\pi sb)]/(2\pi sb)^3 \,. \tag{5.83}$$

Although simple, the resulting expression for the rotation function has the disadvantage of containing intricate $\mathbf{h}$, $\mathbf{k}$ and $[C]^{-1}$ contributions, which makes its calculation time-consuming if the whole range of rotations must be explored. Crowther (1972) showed the advantage of expanding the Patterson functions within a spherical region in terms of spherical harmonics. Indeed, β-sections of the rotation function are calculated with two-dimensional fast Fourier transforms. Further improvements of these fast rotations functions have been made by Navaza (1993) and used in his molecular replacement computer program AMORE (Navaza, 1994). The detailed complicated mathematics has been described by Crowther (1972) and Navaza (1993).

If an asymmetric unit contains more than one copy of a molecule, the rotation matrix between the molecules can be determined by a self-rotation function. Here, the crystal Patterson is rotated against itself and the integration is taken over the volume U around the origin in the same manner. The identical molecules may have an arbitrary orientation to each other, or they may be related by local or so-called non-crystallographic symmetries. Searching for local rotation axes is best carried out in a polar angle system. The search Patterson is brought into each polar orientation and then rotated around the angle value for the local axis being sought, for example, 120° for a threefold local axis.

The rotation function possesses symmetries, which depend on the symmetries of the actual Patterson functions. Methods to determine the asymmetric unit of the relevant rotation function have been developed by Tollin and Rossmann (1966), Narasinga Rao et al. (1980), and Moss (1985). The calculations must be made for this asymmetric unit only, and the computing time is shortened considerably, especially if higher symmetries are present.

A practical point of view is the use of an optimal resolution range. Low-resolution data can be excluded because they are rather insensitive to rotation and contain a considerable contribution of the solvent. High-resolution data can be used in self-rotation functions, but should be excluded for a search model because they are more sensitive for the model. A range between 3 and 5 Å has often proved to be optimal.

5.5.2
Locked Rotation Function

One can use the rotational NCS, determined by the self-rotation function, to increase the signal-to-noise ratio of cross-rotation functions (Tong and Rossmann, 1990). If $\{[S]_n, n = 1, \ldots, N\}$ constitutes the set of NCS rotations, including the identity, and $[C]$ is a correct orientation of the cross-rotation, then $[S]_n[C]$ must also correspond to a correct orientation. Here, we are assuming that the rotational NCS is represented by a group. Now, a function may be defined, the locked cross-rotation, whose values are the average of the cross-rotation values at orientations related by the rotational NCS:

$$R_{LC}([C]) = \sum_{n=1}^{N} R([S]_n[C])/N \; . \tag{5.84}$$

If we substitute the target Patterson function by the average over the NCS of the rotated target functions we can compute R_{LC} as an ordinary cross-rotation, similar to Eq. (5.80).

$$\begin{aligned} R_{LC}([C]) &= \sum_{n=1}^{N} (1/V) \int_U P_t(\mathbf{u}) P_s([C]^{-1}[S]_n^{-1}\mathbf{u}) \mathrm{d}^3\mathbf{u}/N \\ &= (1/V) \int_U \left[\sum_{n=1}^{N} P_t([S]_n\mathbf{u})/N \right] P_s([C]^{-1}\mathbf{u}) \mathrm{d}^3\mathbf{u} \end{aligned} \tag{5.85}$$

In certain cases the NCS can be defined in advance, and this can be used for the calculation of a locked self-rotation function. The NCS is given with elements $\{[I]_n, n = 1, \ldots, N\}$ in a reference orientation. They are connected with the actual NCS elements by

$$[S]_n = [C]_n[I]_n[C]_n^{-1} \; . \tag{5.86}$$

Since each $[S]_n$ should be related to a local maximum of the self-rotation, the function

$$R_{LS}([C]) = \sum_{n=1}^{N} R([C]_n[I]_n[C]_n^{-1})/N \tag{5.87}$$

should also display a maximum for each $[C]_n$, but will have a noise level reduced by $N^{1/2}$.

5.5.3 Translation Function

Once the orientation of the molecule(s) has been determined, translation of the molecule(s) can be obtained from translation searches. These may be carried out in terms of crystallographic R factor, the correlation coefficient, Patterson correlation, and electron-density correlation criteria. As the orientation of the search model is known, the structure factors can be calculated as Eq. (5.88) (Rae, 1977)

$$\mathbf{F}_{\text{calc}}(\mathbf{h}) = \sum_{m} \mathbf{F}_m(\mathbf{h}) \exp\{2\pi i \mathbf{h}[T_m]\mathbf{x}_0\} \ , \tag{5.88}$$

where $\mathbf{x}_0$ is a translation vector from a reference position for the search model. The summation goes over the crystallographic symmetry operators of the space group, $[T_m], \mathbf{t}_m$, which relates the coordinate vectors $\mathbf{x}_{1,j}$ in the first asymmetric unit with the coordinate vectors $\mathbf{x}_{m,j}$ according to

$$\mathbf{x}_{m,j} = [T_m]\mathbf{x}_{1,j} + \mathbf{t}_m \ . \tag{5.89}$$

$$\mathbf{F}_m(\mathbf{h}) = \sum_{j} f_j \exp\{2\pi i \mathbf{h}([T_m]\mathbf{x}_{1,j} + \mathbf{t}_m)\} \tag{5.90}$$

is the contribution of the search model in the mth crystallographic asymmetric unit to the structure factor at the reference position.

5.5.3.1 R-Factor and Correlation-Coefficient Translation Functions

The crystallographic R-factor is often used as criterion in a translation search. It is defined as the difference between the observed ($F_{\text{obs}}(\mathbf{h})$) and calculated ($F_{\text{calc}}(\mathbf{h})$) structure factor amplitudes (Eq. (5.91),

$$R_F = \sum_{\mathbf{h}} |F_{\text{obs}}(\mathbf{h}) - k_F F_{\text{calc}}(\mathbf{h})| / \sum_{\mathbf{h}} F_{\text{obs}}(\mathbf{h}) \ . \tag{5.91}$$

Another approach is to calculate a correlation factor, which has the advantage of being independent of a scale factor like k_F in Eq. (5.91). It is defined as:

$$cc_F = \frac{\sum_{\mathbf{h}}(F_{\text{obs}}(\mathbf{h}) - \langle F_{\text{obs}}(\mathbf{h})\rangle)(F_{\text{calc}}(\mathbf{h}) - \langle F_{\text{calc}}(\mathbf{h})\rangle)}{\left[\sum_{\mathbf{h}}(F_{\text{obs}}(\mathbf{h}) - \langle F_{\text{obs}}(\mathbf{h})\rangle)^2(F_{\text{calc}}(\mathbf{h}) - \langle F_{\text{calc}}(\mathbf{h})\rangle)^2\right]^{1/2}}$$
$$= \frac{\sum_{\mathbf{h}} F_{\text{obs}}(\mathbf{h})F_{\text{calc}}(\mathbf{h}) - \sum_{\mathbf{h}} F_{\text{obs}}(\mathbf{h}) \sum_{\mathbf{h}} F_{\text{calc}}(\mathbf{h})/N}{\left\{\left[\sum_{\mathbf{h}}(F_{\text{obs}}(\mathbf{h}))^2 - \left(\sum_{\mathbf{h}} F_{\text{obs}}(\mathbf{h})\right)^2/N\right]\left[\sum_{\mathbf{h}}(F_{\text{calc}}(\mathbf{h}))^2 - \left(\sum_{\mathbf{h}} F_{\text{calc}}(\mathbf{h})\right)^2/N\right]\right\}^{1/2}} \tag{5.92}$$

Analogical factors can be formulated on the basis of intensities. The structure factor calculation can be split up into two steps. The $\mathbf{F}_m(\mathbf{h})$ values given in Eq. (5.90) are constant for a given orientation of the search model, and can be calculated by placing the search model in a $P1$ unit cell with the same cell dimensions as the unknown crystal unit cell. The calculation of the structure factors according to Eq. (5.88) is then straightforward.

If more than one molecule is present in the asymmetric unit the calculation has to start with one molecule and its translation hopefully found by the aid of these search criteria. It must be stressed, however, that these factors are very sensitive to the completeness of the search model, and it is evident that in the easiest case of two molecules in the asymmetric unit the computation of factors includes 50% of the search model at best. In the lucky case that the first molecule could be positioned this will be kept fixed at its correct position and included into a new translation search for the next molecule. This procedure will be repeated as far as all molecules have been located.

5.5.3.2 Patterson-Correlation Translation Function

In analogy to the rotation functions, a Patterson-correlation function can also be defined for a translation search (e.g., see Tong, 1993). Rotation functions are based on the overlap of only a subset of the interatomic vectors in the Patterson map, the self-vectors within each crystallographically unique molecule. The correct orientation and position of a molecule in the crystal unit cell should lead to the maximum overlap of both the self- and cross-vectors – that is, maximum overlap between the observed (target) and calculated Patterson functions over the whole unit cell. The definition of the Patterson-correlation translation function is then given as:

$$PC(\mathbf{x}_0) = \int_U P_t(\mathbf{u})P_{\text{calc}}(\mathbf{u}, \mathbf{x}_0)\mathrm{d}\mathbf{u} = \sum_{\mathbf{h}}(\mathbf{F}_{\text{obs}}(\mathbf{h}))^2(\mathbf{F}_{\text{calc}}(\mathbf{h}, \mathbf{x}_0))^2 , \tag{5.93}$$

where the integration goes over the whole unit cell.

Substituting Eq. (5.88), we obtain Eq. (5.94):

$$PC(\mathbf{x}_0) = \sum_{\mathbf{h}} \sum_{m} \sum_{n} (F_{\text{obs}}(\mathbf{h}))^2 \mathbf{F}_m(\mathbf{h}) \mathbf{F}_{\mathbf{n}}^*(\mathbf{h}) \times \exp\{-2\pi i \mathbf{h}([T_n] - [T_m])\mathbf{x}_0\} \; . \tag{5.94}$$

This equation is similar to that introduced by Crowther and Blow (1967).

As the Patterson-correlation translation function is computed on an arbitrary scale, it is difficult to compare results from different calculations. The *R*-factor or the correlation coefficient can be calculated for the top peaks of the Patterson-correlation translation function (Eq. (5.94) to place the results on an absolute scale, but other normalization methods may also be used (see Tong, 2001).

5.5.3.3 **Phased Translation Function**

If an atomic model needs to be placed in an electron density map that has been obtained from other sources (e.g., MIR or partial model phases), an electron density correlation translation function (or to use its common name, a phased translation function) (Read and Schierbeek, 1988; Bentley and Houdusse, 1992; Tong, 1993) can be defined:

$$PTF(\mathbf{x}_0) = \sum_{\mathbf{h}} \mathbf{F}_{\text{obs}}(\mathbf{h}) \mathbf{F}_{\text{calc}}^*(\mathbf{h}) = \sum_{\mathbf{h}} \sum_{m} \mathbf{F}_{\text{obs}}(\mathbf{h}) \mathbf{F}_m^*(\mathbf{h}) \exp\{-2\pi i \mathbf{h}[T_m]\mathbf{x}_0\} \; . \tag{5.95}$$

As with the Patterson-correlation translation function, the phased translation function can be put on an absolute scale by introducing appropriate normalization factors, or by converting the results to *R*-factors or correlation coefficients. It has to be considered that the initially used phases could be in the wrong hand, so that the enantiomorph phases must also be tried in the phased translation function.

The stationary molecules contribute a constant to the phased translation function, and this is not shown in Eq. (5.95). However, the phase information from the stationary molecules can be applied to the observed structure factor amplitudes, and the phased translation function, rather than the Patterson-correlation function, can be used in the search for additional molecules (Read and Schierbeek, 1988; Bentley and Houdusse, 1992). This may be especially useful in locating the last few molecules in cases where there are several molecules in the asymmetric unit.

The region of the unit cell that should be covered during a translation search does not generally correspond to the asymmetric unit of the space group. Since the search model has a defined orientation, it can only reside in one of the asymmetric units in the unit cell. Lacking knowledge as to which asymmetric unit the model occupies, the whole unit cell would need to be searched. Once the first molecule is positioned, the origin of the unit cell is also fixed. The search for the subsequent molecules will need to cover the whole cell.

If NCS is present and has been determined before, it can be used to compute a locked translation function (Tong, 1996). It can determine the position of the

monomer search model relative to the center of the NCS assembly. Using this information, the whole assembly can be generated and can then be used in a conventional translation search to locate the center of this NCS assembly in the unit cell.

A packing check in translation search may be used to exclude unreasonable translation solutions.

5.5.4 Computer Programs for Molecular Replacement

An early program for molecular replacement, working in direct space, was written by Huber (1965). Today, several program packages are available, either being dedicated exclusively to the molecular replacement technique, or having integrated relevant modules. Pure molecular replacement programs include AMORE (Navaza, 1994) and GLRF (Tong and Rossmann, 1990), MOLREP (Vagin and Teplyakov, 1997) and BEAST (Read, 2001). The rotational and translational search starting from the search model is fully automated in AMORE, and includes a final rigid body refinement of each proposed solution. GLRF offers different types of rotation and translation functions, all operating in reciprocal space, and a Patterson correlation refinement (Brünger, 1990). One peculiarity of the GLRF program is the locked rotation function, which takes into account possible NCS symmetries and is an average of *n* independent rotation functions with an improved peak-to-noise ratio. BEAST uses likelihood-based molecular replacement methods. Other frequently used program packages, including molecular replacement modules, are the CCP4 program suite (CCP4, 1994), CNS (Brünger et al., 1998) and PROTEIN (Steigemann, 1991).

References

Bentley, G.A., Houdusse, A., *Acta Crystallogr.* **1992**, *A48*, 312–322.

Bhuiya, A.K., Stanley, E., *Acta Crystallogr.* **1963**, *16*, 981–984.

Blessing, R.H., *J. Appl. Crystallogr.* **1997**, *30*, 176–177.

Blessing, R.H., Guo, D.Y., Langs, D.A., *Acta Crystallogr.* **1996**, *D52*, 257–266.

Blessing, R.H., Smith, G.D., *J. Appl. Crystallogr.* **1999**, *32*, 664–670.

Blow, D.M., Crick, F.H.C., *Acta Crystallogr.* **1959**, *12*, 794–802.

Brünger, A.T., *Acta Crystallogr.* **1990**, *A46*, 46–57.

Brünger, A.T., Adams, P.D., Clore, G.M., Delano, W.L., Gros, P., Grossekunstleve, R.W., et al., *Acta Crystallogr.* **1998**, *D54*, 905–921.

Buerger, M.J., *Vector Space*, John Wiley, New York, **1959**.

Carvin, D.G.A., Islam, S.A., Sternberg, M.J.E., Blundell, T.L., *Isomorphous Replacement and Anomalous Scattering*. Daresbury Laboratory, Warrington, **1991**.

CCP4, *Acta Crystallogr.* **1994**, *D50*, 760–763.

Cochran, W., *Acta Crystallogr.* **1955**, *8*, 473–478.

Crick, F.H.C., Magdoff, B.S., *Acta Crystallogr.* **1956**, *9*, 901–908.

Cromer, D.T., Liberman, D., *J. Chem. Phys.* **1970**, *53*, 1891–1898.

Crowther, R.A., The fast rotation function. In: Rossmann, M.G. (Ed.), *The Molecular*

Replacement Method. Gordon and Breach, New York, **1972**.
Crowther, R.A., Blow, D.M., *Acta Crystallogr.* **1967**, *23*, 544–548.
Debaerdemaker, T., Woolfson, M.M., *Acta Crystallogr.* **1983**, *A39*, 193–196.
Dickerson, E.E., Weinzierl, J.E., Palmer, R.A., *Acta Crystallogr.* **1968**, *B24*, 997–1003.
Dodson, E., Vijayan, M., *Acta Crystallogr.* **1971**, *B27*, 2402–2411.
Drenth, J., *Principles of Protein X-ray Crystallography*. Springer, New York, **1994**.
Germain, G., Woolfson, M.M., *Acta Crystallogr.* **1968**, *B24*, 91–96.
Green, D.W., Ingram, V.M., Perutz, M.F., *Proc. R. Soc. London Ser. A* **1954**, *225*, 287–307.
Hauptman, H.A., A minimal principle in the phase problem. In: Moras, D., Podjarny, A.D., Thierry, J.C. (Eds.), *Crystallographic Computing* 5: *from Chemistry to Biology*, pp. 324–332. International Union of Crystallography and Oxford University Press, Oxford, **1991**.
Hauptman, H.A., Karle, J., *Solution of the Phase Problem, I. The Centrosymmetric Crystal*. Americ. Crystallogr. Assoc. Monograph No. 3, Polycrystal Book Service, Dayton, **1953**.
Hendrickson, W.A., *Science* **1991**, *254*, 51–58.
Hendrickson, W.A., Horton, J.R., LeMaster, D.M., *EMBO J.* **1990**, *9*, 1665–1672.
Hendrickson, W.A., Ogata, C.M., *Methods Enzymol.* **1997**, *276*, 494–523.
Hönl, H., *Z. Phys.* **1933**, *84*, 1–16.
Hoppe, W., *Z. Elektrochem.* **1957**, *61*, 1076–1083.
Huber, R., *Acta Crystallogr.* **1965**, *19*, 353–356.
Islam, S.A., Carvin, D., Sternberg, M.J.E., Blundell, T.L., *Acta Crystallogr.* **1998**, *D54*, 1199–1206.
James, R.W., The optical principles of the diffraction of X-rays. In: *The Crystalline State*, Vol. II. pp. 135–167. Bell, London, **1960**.
Karle, J., Hauptman, H., *Acta Crystallogr.* **1956**, *9*, 635–651.
Knäblein, J., Neuefeind, T., Schneider, F., Bergner, A., Messerschmidt, A., Löwe, J., Steipe, B., Huber, R., *J. Mol. Biol.* **1997**, *270*, 1–7.
Knight, S.D., *Acta Crystallogr. Section D* **2000**, *56*, 42–47.
de La Fortelle, E., Bricogne, G., *Methods Enzymol.* **1997**, *276*, 472–494.
Main, P., Fiske, S.J., Hull, S.E., Lessinger, L., Germain, G., Declercq, J.-P., Woolfson, M.M., MULTAN 80: A System of Computer Programs for the Automatic Solution of Crystal Structures from X-ray Diffraction Data. Universities of York and Louvain, **1980**.
Matthews, B.W., Czerwinski, E.W., *Acta Crystallogr.* **1975**, *A31*, 480–497.
Miller, R., Gallo, S.M., Khalak, H.G., Weeks, C.M., *J. Appl. Crystallogr.* **1994**, *27*, 613–621.
Moss, D.S., *Acta Crystallogr.* **1985**, *A41*, 470–475.
Narasinga Rao, S., Jyh-Hwang Jih, Hartsuck, J.A., *Acta Crystallogr.* **1980**, *A36*, 878–884.
Navaza, J., *Acta Crystallogr.* **1993**, *D49*, 588–591.
Navaza, J., *Acta Crystallogr.* **1994**, *A50*, 157–163.
Nordman, C.E., *Trans. Am. Crystallogr. Assoc.* **1966**, *2*, 29–38.
Otwinowski, Z., Maximum likelihood refinement of heavy atom parameters. In: Wolf, W., Evans, P.R., Leslie, A.G.W. (Eds.), *Isomorphous Replacement and Anomalous Scattering*, pp. 80–86. SERC Daresbury Laboratory, Warrington, **1991**.
Pearson, R.G., *Survey* **1969**, *5*, 1–52.
Rae, A.D., *Acta Crystallogr.* **1977**, *A33*, 423–425.
Read, R.J., *Acta Crystallogr.* **2001**, *D57*, 1373–1382.
Read, R.J., Schierbeek, A.J., *J. Appl. Crystallogr.* **1988**, *21*, 490–495.
Rice, S.O., *Bell System Tech. J.* **1944**, *23*, 283–332.
Rossmann, M.G., *Acta Crystallogr.* **1960**, *13*, 221–226.
Rossmann, M.G., Blow, D.M., *Acta Crystallogr.* **1962**, *15*, 24–31.
Sheldrick, G.M., *Acta Crystallogr.* **1990**, *A46*, 467–473.
Sheldrick, G.M., Heavy atom location using SHELXS-90. In: Wolf, W., Evans, P.R., Leslie, A.G.W. (Eds.), *Proceedings of the*

CCP4 Study Weekend. Isomorphous Replacement and Anomalous Scattering, pp. 23–38. Daresbury Laboratory, Warrington, **1991**.

Sheldrick, G.M., Gould, R.O., *Acta Crystallogr.* **1995**, *B51*, 423–431.

Sheldrick, G.M., Direct methods based on real/reciprocal space iteration. In: Wilson, K.S., Davies, G., Ashton, A.S., Bailey, S. (Eds.), *Proceedings of the CCP4 Study Weekend. Recent Advances in Phasing*, pp. 147–158. Daresbury Laboratory, Warrington, **1997**.

Sheldrick, G.M., SHELX: Application to macromolecules. In: Fortier, S. (Ed.), *Direct Methods for Solving Macromolecular Structures*, pp. 401–411. Kluwer Academic Publishers, Dordrecht, **1998**.

Sheldrick, G.M., Hauptman, H.A., Weeks, C.M., Miller, R., Usón, Ab initio phasing. In: Rossmann, M.G., Arnold, E. (Eds.), *International Tables for Crystallography*, Vol. F, pp. 333–345. Kluwer Academic Publishers, Dordrecht, **2001**.

Shmueli, U., Wilson, A.J.C., Statistical properties of the weighted reciprocal lattice. In: Shmueli, U. (Ed.), *International Tables for Crystallography, Vol. B, Reciprocal Space*. Kluwer Academic Publishers, Dordrecht, **2001**.

Steigemann, W., Recent advances in the PROTEIN program system for the X-ray structure analysis of biological macromolecules. In: Moras, D., Podjarny, A.D., Thierry, J.C. (Eds.), *Crystallographic Computing 5: from Chemistry to Biology*, pp. 115–125. Oxford University Press, Oxford, **1991**.

Terwilliger, T.C., Kim, S.-H., Eisenberg, D., *Acta Crystallogr.* **1987**, *A43*, 1–5.

Terwilliger, T.C., Berendzen, J., *Acta Crystallogr.* **1999**, *D55*, 849–861.

Tollin, P., Rossmann, M.G., *Acta Crystallogr.* **1966**, *21*, 872–876.

Tong, L., *J. Appl. Crystallogr.* **1993**, *26*, 748–751.

Tong, L., *Acta Crystallogr.* **1996**, *A52*, 476–479.

Tong, L., Translation functions. In: Rossmann, M.G., Arnold, E. (Eds.), *International Tables for Crystallography*, Vol. F. Kluwer Academic Publishers, Dordrecht, **2001**.

Tong, L., Rossmann, M.G., *Acta Crystallogr.* **1990**, *A46*, 783–792.

Vagin, A.A., Teplyakov, A., *J. Appl. Crystallogr.* **1997**, *30*, 1022–1025.

Weeks, C.M., DeTitta, G.T., Miller, R., Hauptmann, H.A., *Acta Crystallogr.* **1993**, *D49*, 179–181.

Weeks, C.M., Hauptman, H.A., Chang, C.-S., Miller, R., *ACA Trans. Symp.* **1994**, *30*, 153–161.

Weeks, C.M., Miller, R., *J. Appl. Crystallogr.* **1999**, *32*, 120–124.

Yang, W., Hendrickson, W.A., Kalman, E.T., Crouch, R.J., *J. Biol. Chem.* **1990**, *265*, 13553–13559.

6
Phase Improvement by Density Modification and Phase Combination

6.1
Introduction

With the methods so far described, an experimental electron density map can be calculated and if its quality is sufficiently high, the atomic model can be constructed. However, there are methods for further phase improvement available which may be applied in general, or depending on given prerequisites. Such phase improvement routines have been used routinely over the past 20 years and have had a large impact on the advancement of biological macromolecular crystallography. One group of these routines uses the technique of density modification, which is based on some conserved features of the correct electron density function. Several features or constraints can be used to improve the quality of the experimental electron density map:

- flatness of the solvent region of the biomacromolecular crystal, used in solvent flattening;
- ideal electron density distribution, used in histogram matching;
- NCS, used in molecular averaging;
- protein backbone connectivity, used in skeletonization;
- local shape of electron density, used with Sayre's equation;
- atomicity, used in atomization of the electron density function;
- structure-factor amplitudes, used in Sim weighting; and
- experimental phases, used in phase combination.

Although this list is not long, it covers the most widely used methods.

In density-modification techniques, the chemical and physical constraints of the electron density function are in real space but must be applied to amplitudes and phases, which are in reciprocal space. Amplitudes and phases have good estimates of error, which is not the case for the constraints in real space. These represent expectations about the structure which may be difficult to quantify, and therefore an iterative method is used with real- and reciprocal-space steps, as shown in Figure 6.1. A weighted map is computed and used as a basis for applying all the real-space modifications. The modified map is then Fourier back-transformed to generate a set of amplitudes and phases. The agreement between the observed amplitudes and the amplitudes calculated from the

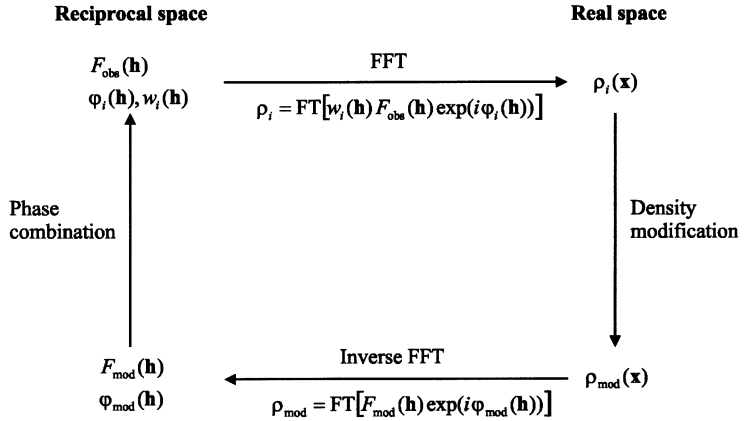


Fig. 6.1 The cycle for iterative application of real-space and reciprocal-space constraints in density-modification methods.

modified map is then used to estimate weights for the modified phases, which in turn are used to combine the modified phases with experimental phases to generate new phases.

6.2 Solvent Flattening

Protein crystals have a solvent content of between 75 and 40%. In a highly refined protein crystal structure, the solvent space between the molecules is rather flat owing to the dynamic nature of this region. Usually, the initial experimental starting phases are of lower quality than the final ones, and as a result the solvent region (if the molecular boundaries can be identified) contains noise peaks. It is now obvious to set the noisy solvent space to a low constant value and to calculate the new improved phases by Fourier back-transforming this corrected electron density map. However, it is evident that the definition of the molecular boundaries will be tedious and depend on the quality of the electron density map. Wang (1985) has proposed an automatic procedure which smoothes the electron density to define the protein region. This smoothed electron density map is traced against a threshold value which separates this map into molecule and solvent space according to their ratio of volumes in the unit cell. The space inside the molecular envelope is polished to avoid internal voids. Then, a new electron density map is calculated using the observed structure factor amplitudes and the phases are revealed from the solvent corrected map. The solvent-corrected map is obtained by setting all electron density values inside the molecular envelope to those of the initial map, and all values outside the envelope to a low constant value. These phases from the solvent-flattening procedure can be combined with the MIR or MAD phases. This procedure can be repeated in sev-

eral iterative cycles because, after each cycle of solvent flattening, the quality of the electron density map is improved. There are no prerequisites for the application of the method of solvent flattening. It is evident that solvent flattening is most effective for crystals with a high solvent content.

The solvent content of the protein crystal is an important input parameter for the solvent-flattening procedure. It can be estimated by virtue of the Matthews parameter V_M, which is defined according to Eq. (6.1):

$$V_M = \frac{V_{\text{unit cell}}}{M_{\text{Prot}}} \tag{6.1}$$

where $V_{\text{unit cell}}$ is the volume of the unit cell and M_{Prot} is the molecular mass of the protein in the unit cell. V_M has values in the range of 1.6 to 3.5 $\text{Å}^3\ \text{Da}^{-1}$ for proteins. This allows, first, a rough estimation to be made of the number of molecules in the unit cell. Furthermore, as mentioned above, V_M can be used to assess the solvent content of a protein crystal. By calling V_{Prot} the crystal volume occupied by the protein, V'_P its fraction with respect to the total crystal volume V and M_{Prot} the mass of protein in the cell, we obtain:

$$V'_P = \frac{V_{\text{Prot}}}{V} = \frac{V_{\text{Prot}}/M_{\text{Prot}}}{M_{\text{Prot}}/V} \tag{6.2}$$

The first term is the specific volume of the protein, the second the reciprocal of V_M and, remembering that the molecular weight is expressed in Daltons, we have:

$$V'_P = \frac{1.6604}{d_{\text{Prot}} V_M} \tag{6.3}$$

Taking 1.35 g cm^{-3} as the protein density, we obtain as first approximation Eqs. (6.4) and (6.5):

$$V'_P \approx \frac{1.23}{V_M} \tag{6.4}$$

$$V'_{\text{Solv}} \approx 1 - V'_P \tag{6.5}$$

The calculation of the smoothed map in the original proposal of Wang (1985) was carried out in real space and was very intensive in computing time. Shortly thereafter, Leslie (1987) found that the calculation could be performed very efficiently in reciprocal space using fast Fourier transforms (FFTs). The map is smoothed by calculating at each point in the map the mean density over the encompassing sphere of radius R. This operation can be written as a convolution of the truncated map ρ_{trunc} with a special weighting function $w(\mathbf{r})$ (Eq. 6.6),

$$\rho_{\text{smooth}}(\mathbf{x}) = \sum_{\mathbf{r}} w(\mathbf{r}) \rho_{\text{trunc}}(\mathbf{x} - \mathbf{r}) \ , \tag{6.6}$$

where (Eq. 6.7)

$$w(\mathbf{r}) = \begin{cases} 1 - |\mathbf{r}|/R, & |\mathbf{r}| \leq R \\ 0, & |\mathbf{r}| \geq R\,. \end{cases} \tag{6.7}$$

The truncated map is calculated according to (Eq. 6.8)

$$\rho_{\text{trunc}}(\mathbf{x}) = \begin{cases} \rho(\mathbf{x}), & \rho(\mathbf{x}) \geq \rho_{\text{solv}} \\ 0, & \rho(\mathbf{x}) \leq \rho_{\text{solv}}\,. \end{cases} \tag{6.8}$$

The convolution of Eq. (6.6) can now be written according to the convolution theorem as (Eq. 6.9)

$$\rho_{\text{smooth}}(\mathbf{x}) = \text{FT}^{-1}\{\text{FT}[\rho_{\text{trunc}}(\mathbf{x})]\text{FT}[w(\mathbf{r})]\}\,, \tag{6.9}$$

where FT denotes a Fourier transform and FT^{-1} represents an inverse Fourier transform. The Fourier transform of the truncated map is readily calculated using standard FFT programs, and it can be shown that the Fourier transform of the weighting function is given by Eq. (6.9)

$$g(S) = \text{FT}[w(\mathbf{r})] = Y(2\pi RS) - Z(2\pi RS)\,, \tag{6.10}$$

where

$$S = 2\sin\theta/\lambda$$

$$Y(2\pi RS) = 3[\sin(2\pi RS) - 2\pi RS\cos(2\pi RS)]/(2\pi RS)^3 \tag{6.11}$$

$$Z(2\pi RS) = 3\{4\pi RS\sin(2\pi RS) - [(2\pi RS)^2 - 2]\cos(\pi RS) - 2\}/(2\pi RS)^4\,. \tag{6.12}$$

Other weighting functions may be implemented by the same approach.

Now, the boundary between the protein molecule and the solvent region can be determined from the smoothed electron density map. A cut-off value ρ_{cut} is calculated, which divides the unit cell into two regions occupying the correct volumes for the protein and the solvent. All points in the map where $\rho_{\text{smooth}}(\mathbf{x}) \leq \rho_{\text{cut}}$ are defined to be in the solvent region, and a molecular mask or envelope is obtained as result of this procedure. The radius of the sphere R for the smoothing of electron densities is generally chosen at around 8 Å and, in most cases, this delivers satisfying results.

The modified electron density map $\rho_{\text{mod}}(\mathbf{x})$ is now gained by setting all points to the original values $\rho(\mathbf{x})$ for $\rho_{\text{smooth}}(\mathbf{x}) \geq \rho_{\text{cut}}$ and to ρ_{solv} for $\rho_{\text{smooth}}(\mathbf{x}) \leq \rho_{\text{cut}}$. ρ_{solv} is the expected value for the solvent region. If the electron density has not been calculated on an absolute scale, the solvent density can be set to its mean value.

In 1996, Abrahams and Leslie improved the method by introducing so-called "solvent flipping". For all grid points within the solvent the electron density is

set to $\rho_{\mathrm{mod}}(\mathbf{x}) = \rho_{\mathrm{solv}} - [\gamma/(1-\gamma)][\rho(\mathbf{x}) - \rho_{\mathrm{solv}}]$ with a relaxation factor γ. This corresponds to a flipping of the features of the solvent and corrects for the problem of independence in phase combination, which will be discussed later.

6.3 Histogram Matching

Histogram matching is a technique emanating from image processing. It is aimed at bringing the density distribution of an image to an ideal distribution, thereby improving the image quality. In the application to electron density maps it is assumed that a high-quality protein crystal structure has a characteristic frequency distribution of electron density which serves as a standard reference distribution for other electron density maps. Such maps of lower quality exhibit a frequency distribution of electron density which deviates from the standard distribution. Zhang and Main (1988) systematically examined the electron density histogram of several proteins, and noted that the ideal density histogram depends on resolution, the overall temperature factor, and the phase error. It is, however, independent of structural conformation. The frequency distribution may be treated as function of resolution only if the overall temperature factor is sharpened to $B_{\mathrm{overall}} = 0$.

Beside the derivation of ideal electron density histograms from known protein structures, such histograms can also be predicted by an analytical formula. We present the method of Main (1990) in more detail, following the elaborations by Zhang et al. (2001). The density histogram is split into components that are related to three types of electron density in the map:

1. A region of overlapping densities, which can be represented by a randomly distributed background noise with a histogram expressed by a Gaussian distribution (Eq. 6.13),

$$P_o(\rho) = N \exp[-(\rho - \bar{\rho})^2/2\sigma^2] \,, \tag{6.13}$$

   where $\bar{\rho}$ is the mean density and σ is the standard deviation.
2. A region of partially overlapping densities with a histogram modeled by a cubic polynomial function (Eq. 6.14),

$$P_{po}(\rho) = N(ap^3 + bp^2 + cp + d) \,. \tag{6.14}$$

3. A region of non-overlapping atomic peaks with a histogram derived analytically from a Gaussian atom (Eq. 6.15),

$$P_{no}(\rho) = N(A/\rho)[\ln(\rho_0/\rho)]^{1/2} \,, \tag{6.15}$$

   where ρ_0 is the maximum density, N is a normalizing factor and A is the relative weight of the terms between Eq. (6.13) and Eq. (6.15). If we use two

threshold values, ρ_1 and ρ_2, regions (1) to (3) are defined as $2\rho \leq \rho_2$, $2\rho_2 < \rho \leq \rho_1$ and $2\rho_1 < \rho \leq \rho_0$, respectively, with the corresponding Eqs. (6.13) to (6.15). The parameters *a, b, c, d* in the cubic polynomial are calculated by matching function values and gradients at ρ_1 and ρ_2. The parameters in the histogram formulae, $\bar{\rho}, \sigma, A, \rho_0, \rho_1, \rho_2$, can be taken from histograms of known structures.

Histogram matching is now performed in altering the experimental map with its imperfect histogram in such a way as to make its density histogram equal to the ideal distribution. The procedure is explained in Figure 6.2. Let $P(\rho)$ be the current density histogram and $P'(\rho)$ the desired distribution, normalized such that their sums are equal to 1. We can then calculate the cumulative distribution functions, $N(\rho)$ and $N'(\rho)$, according to Eqs. (6.16) and (6.17):

$$N(\rho) = \int_{\rho_{\min}}^{\rho} P(\rho)\mathrm{d}\rho \ , \tag{6.16}$$

$$N'(\rho') = \int_{\rho_{\min}}^{\rho} P'(\rho)\mathrm{d}\rho \tag{6.17}$$

The cumulative distribution function of a variable transforms a value taken from the distribution into a number between 0 and 1, representing the position of that value in an ordered list of values taken from the distribution.

The transformation is now made in the following way: A density value of $P(\rho)$ (lower left diagram in Fig. 6.2) is chosen, its corresponding value $N(\rho)$ (upper left diagram in Fig. 6.2) is determined, mapped then to the desired cumulative distribution value $N'(\rho)$ (right upper diagram in Fig. 6.2) and the desired modified value ρ' is finally obtained by Eq. (6.18):

$$\rho' = N'^{-1}[N(\rho)] \ . \tag{6.18}$$

The distribution of ρ' will then match the desired distribution after the above transformation. In practice, the density is divided in bins ranging from 1 to *n*, and the transformation in Eq. (6.18) can be made through a linear transform according to Eq. (6.19):

$$\rho'_i = a_i\rho_i + b_i \ . \tag{6.19}$$

The density histogram contains some useful properties of the electron density, such as the minimum, maximum and mean density, the density variance, and the entropy of the map. The latter three magnitudes for the ideal map are obtained as follows:

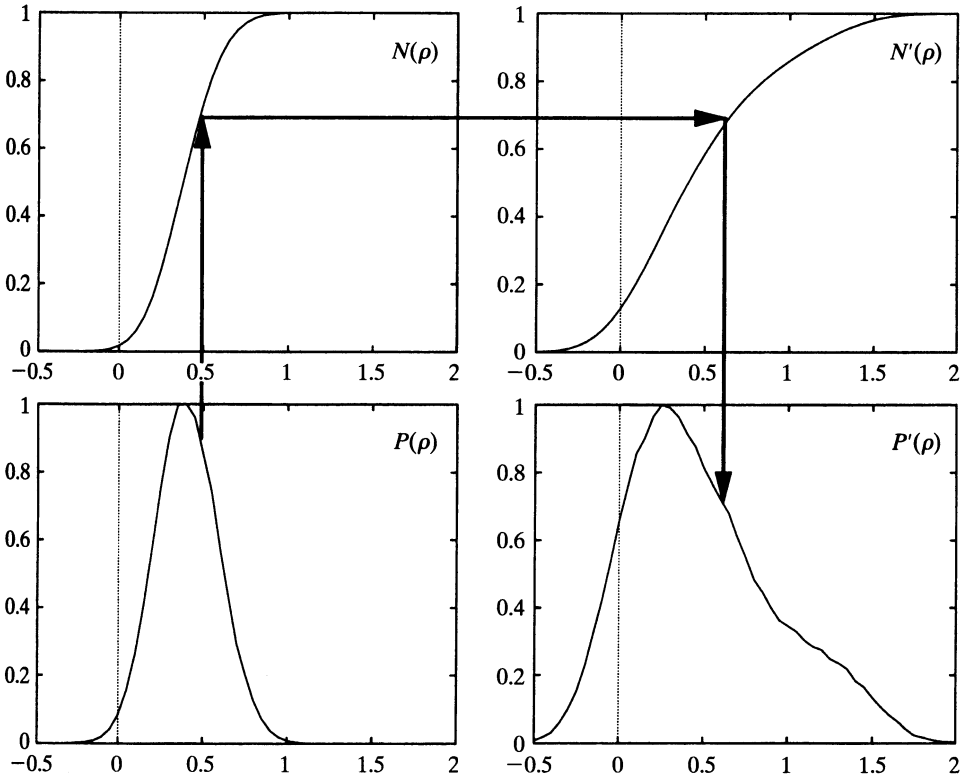


Fig. 6.2 Illustration of the process of histogram matching. (Reproduced by permission of The International Union for Crystallography, from Zhang et al., 2001.)

- mean density $\bar{\rho}$ (Eq. 6.20),

$$\bar{\rho} = \int_{\rho_{\min}}^{\rho_{\max}} \rho P(\rho) \mathrm{d}(\rho) \,, \tag{6.20}$$

- variance of the density $\sigma(\rho)$ (Eq. 6.21)

$$\sigma(\rho) = \left(\overline{\rho^2} - \bar{\rho}^2\right)^{1/2} \,, \tag{6.21}$$

where

$$\overline{\rho^2} = \int_{\rho_{\min}}^{\rho_{\max}} \rho^2 P(\rho) \mathrm{d}\rho \,, \tag{6.22}$$

- entropy S, (Eq. 6.23)

$$S = -\int_{\rho_{min}}^{\rho_{max}} P(\rho)\rho \ln(\rho) d\rho \, . \tag{6.23}$$

Summarizing, the process of histogram matching applies a minimum and a maximum value to the electron density, imposes the correct mean and variance, and defines the entropy of the new map. The rank of electron density values remains unchanged in the modified map. Histogram matching is applied to the protein region and is therefore complementary to solvent flattening, which only operates in the solvent region.

In the process of density modification, structure factors or electron density from different sources are compared and combined. It is, therefore, very important to put all the maps and structure factor amplitudes on a common scale. The density histogram can be used to scale observed structure factor amplitudes (Cowtan and Main, 1993; Zhang, 1993). This is a robust method, which also works well with medium- to low-resolution data, where scaling using Wilson statistics is often inaccurate.

Histogram matching is normally used together with solvent flattening and is incorporated into the density modification programs SQUASH (Zhang and Main, 1990; Zhang, 1993) and DM from the CCP4 program package (CCP4, 1994).

6.4 Molecular Averaging

If there is more than one identical subunit in the asymmetric unit of the crystal, then molecular averaging can be used to improve the protein phases. The spatial relations between the single identical subunits in the asymmetric unit may be determined by Patterson search methods (as described in Section 5.5), or from the arrangement of the heavy atoms or anomalous scatterers. The spatial relation between the identical subunits can be either improper (the relevant spatial movement consists of a rotation about an unsymmetrical angle value and a translation component; Fig. 6.3 a) or proper (the spatial movements form a symmetry group which is composed of rotational symmetry elements only; Fig. 6.3 b). Such additional symmetries are called noncrystallographic (NCS) or local, and there are no limitations concerning the rotational periodicity of the symmetry axes (e.g., five-, seven- and higher-fold axes are allowed). It is evident that averaging about the different related subunits, the electron density of which should be equal in each subunit, must result in an improved electron density map and therefore in improved protein phases. Molecular averaging is best done in direct space, and several programs are available for this, including AVE, RAVE (Kleywegt and Jones, 1994) and MAIN (Turk, 1992).

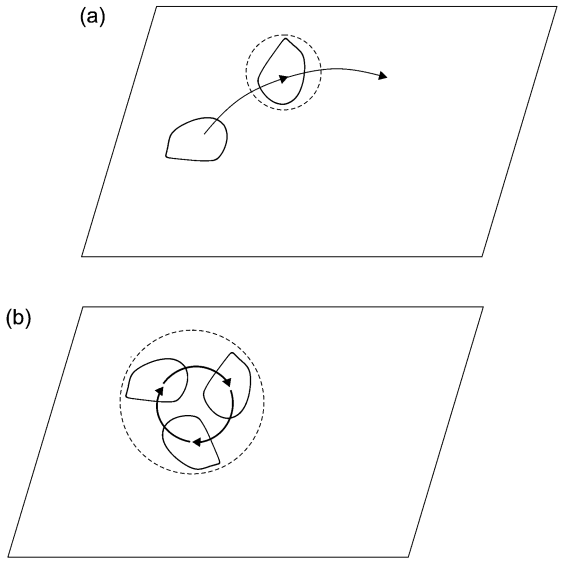


Fig. 6.3 (a) Improper NCS; (b) proper NCS.

The procedure of molecular averaging is composed of several steps:

- First, the molecular envelope must be determined from the initial electron density map or from a molecular model which, for example, has been obtained from molecular replacement. In the case of a molecular model from molecular replacement, this is straightforward. If one starts from an initial electron density map, the mask can be determined using the local density correlation function as developed by Vellieux et al. (1995). The local correlation function distinguishes those volumes of the crystal which map onto similar density under transformations by the NCS operator. Thus, in the case of improper NCS, a local correlation mask will cover only one monomer. In the case of a proper symmetry, a local correlation mask will cover the whole complex (Fig. 6.3 a, b).
- Next, the particular electron density averaging between the related subunits is performed. All grid points in the molecular envelope are passed through, and the respective NCS-related electron density values are mapped to the actual grid point and averaged. The situation for a monomer and a NCS multimer envelope is shown in Figure 6.4 a and b, respectively. In the case of the NCS multimer envelope, the NCS symmetry operators are defined with respect to the center of gravity of the whole multimer envelope. Furthermore, it may be difficult to define the individual NCS monomer. This is best done in the stage of building the atomic model, where connectivities in the polypeptide chain of the protein molecule help to identify the tertiary structure of the molecule. In order to carry out averaging on the basis of the NCS monomer envelope, the NCS operators related to this envelope must be determined, which can be

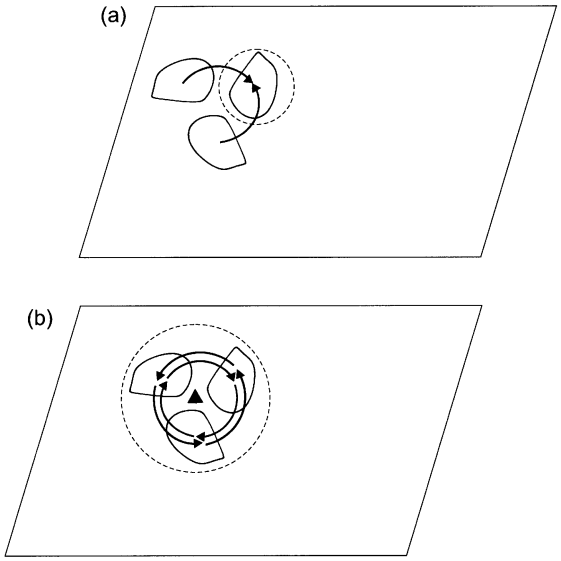


Fig. 6.4 Mapping of NCS-related electron density points onto the molecular envelope. (a) NCS monomer envelope; (b) NCS multimer envelope.

done, for example, in the CNS program system. Since NCS operators will not normally map grid points onto each other, an interpolation of electron density values at non-map grid sites is usually needed. Suitable interpolation functions are described in Bricogne (1974) and Cowtan and Main (1998).

- This is followed by the reconstitution of the complete crystal unit cell with the averaged electron density. The space outside the molecular envelope is flattened, and this map is then Fourier back-transformed. The obtained phase angles can either be taken directly or combined with known phase information to calculate a new and improved electron density map. This cycle can be repeated several times until convergence of the electron density map improvement has been reached. It is very useful to refine the local symmetry operations after every macrocycle of molecular averaging. This can be done with the program IMPROVE, from the Uppsala averaging program collection (Kleywegt and Jones, 1994). Furthermore, molecular averaging can be applied if proteins crystallize in more than one crystal form.

Molecular averaging is especially efficient if a high NCS is present (as in virus structures), but the averaging over two related subunits alone (the lowest case of local symmetry) can give a considerable improvement. In special cases where high NCS symmetry exists and the phase information extends only to low resolution, cyclic molecular averaging can be used to extend the phase angles to provide full resolution of the native protein. This was first shown in the structure analysis of hemocyanin from *Panulirus interruptus* (Gaykema et al., 1984,

1985). It has also been used extensively in the analysis of icosahedral structures (e.g., see Rossmann et al., 1985; Ladenstein et al., 1988) and for large molecular assemblies (e.g., Löwe et al., 1995).

6.5 Sayre's Equation

In Section 5.3.2 we introduced the tangent formula (Eq. 5.57) which calculates the phase for structure factor amplitude $F(\mathbf{h})$ from phases of structure factor pairs $\mathbf{F}(\mathbf{k})$ and $\mathbf{F}(\mathbf{h}-\mathbf{k})$. A related equation has been derived by Sayre (1952) which links the corresponding structure factors directly in amplitude and phase. In reciprocal space, it is given as:

$$\mathbf{F}(\mathbf{h}) = [\theta(\mathbf{h})/V] \sum_{\mathbf{k}} \mathbf{F}(\mathbf{k})\mathbf{F}(\mathbf{h}-\mathbf{k}) \,, \tag{6.24}$$

where $\theta(\mathbf{h})$ is the ratio of scattering factors of real, $f(\mathbf{h})$, and "squared", $g(\mathbf{h})$, atoms and V is the unit cell volume, i.e., Eq. (6.25):

$$\theta(\mathbf{h}) = f(\mathbf{h})/g(\mathbf{h}) \,. \tag{6.25}$$

Sayre's equation is exact for an equal atom structure at atomic resolution. The reciprocal-space function $\theta(\mathbf{h})$ can be calculated according to Eq. (6.25), where both scattering factors can be represented by a Gaussian function. At infinite resolution, one expects $\theta(\mathbf{h})$ to be a spherically symmetric function that decreases smoothly with increased $\mathbf{h}$. However, for data at non-atomic resolution, the $\theta(\mathbf{h})$ curve will behave differently because atomic overlap changes the peak shape. Therefore, a spherical-averaging method is adopted to receive an estimate of the shape function empirically from the ratio of the observed structure factors and the structure factors from the squared electron density by Eq. (6.26)

$$\theta(S) = V \left\langle \mathbf{F}(\mathbf{h}) / \sum_{\mathbf{k}} \mathbf{F}(\mathbf{k})\mathbf{F}(\mathbf{h}-\mathbf{k}) \right\rangle_{|\mathbf{h}|} \,, \tag{6.26}$$

where the averaging is done over the ranges of $|\mathbf{h}|$ – that is, over spherical shells, each covering a narrow resolution range. Here, S denotes the modulus of $\mathbf{h}$.

The empirically derived shape function only extends to the resolution of the experimentally observed phases. This is sufficient for phase refinement, but not for phase extension where no experimentally observed phases are available to provide the empirical $\theta(S)$ for the extension region. Therefore, a Gaussian function of the form

$$\theta(S) = K \exp(-BS^2) \tag{6.27}$$

is fitted to the available values of $\theta(S)$ and parameters K and B are calculated using a least-squares method. The shape function $\theta(S)$ for the resolution beyond that of the observed phases is extrapolated using the fitted Gaussian function.

As Sayre's equation links all structure factor amplitudes and phases, it is a powerful tool for phase refinement and extension. It works perfectly at atomic resolution, but at lower resolution the shape functions $\theta(S)$ of Eqs. (6.26) and (6.27) must be used. It transpired that the derivation of the shape function $\theta(S)$ from a combination of spherical averaging and Gaussian extrapolation was key to the successful application of Sayre's equation for phase improvement at non-atomic resolution (Zhang and Main, 1990).

6.6 Atomization

The atomization method uses the fact that the structure underlying the electron density map consists of discrete atoms. It attempts to interpret the map by automatically placing atoms and refining their positions. A successful application of this method needs atomic resolution of the diffraction data. For non-atomic resolution, Agarwal and Isaacs (1977) proposed a method for the extension of phases to higher resolutions by interpreting an electron density map in terms of "dummy" atoms. The placement of such "dummy" atoms is subject to constraints of bonding distance and the number of neighbors. This procedure was not automated, however, and Lamzin and Wilson (1997) subsequently extended the approach in the ARP (Automated Refinement Program) program for biomacromolecular applications. This technique has become very effective for the solution of structures at high resolution from a poor molecular replacement model, or even directly from an MIR/MAD map.

Perrakis et al. (1997) later developed the wARP program, which performs improvement and extension of crystallographic phases by weighted averaging of multiple-refined dummy atomic models. Map improvement has been demonstrated for maps at medium resolution.

6.7 Phase Combination

During the course of a crystal structure analysis of a biological macromolecule, phase information from different sources may be available, such as information from isomorphous replacement, anomalous scattering, partial structures, solvent flattening, and molecular averaging. An overall phase improvement can be expected when these factors are combined, and a useful method to do this was proposed by Hendrickson and Lattman (1970). The probability curve for each reflection is written in an exponential form as Eq. (6.28):

$$P_s(a) = N_s \exp(K_s + A_s \cos a + B_s \sin a + C_s \cos 2a + D_s \sin 2a) \tag{6.28}$$

Subscript s represents the source from which the phase information has been derived. K_s and the coefficients A_s, B_s, C_s and D_s depend on the structure factor amplitudes and other magnitudes, for example the estimated standard deviation of the errors in the derivative intensity, but are independent of the protein phase angles a. The overall probability function $P(a)$ is obtained by multiplication of the individual phase probabilities, and this turns out to be a simple addition of all K_s and of the related coefficients in the exponential term. Hence, we obtain Eq. (6.29):

$$P(a) = \prod_s P_s(a) = N' \exp\left[\sum_s K_s + \left(\sum_s A_s\right) \cos a + \left(\sum_s B_s\right) \sin a + \left(\sum_s C_s\right) \cos 2a + \left(\sum_s D_s\right) \sin 2a\right] \tag{6.29}$$

K_s and the coefficients A_s–D_s have special expressions for each source of phase information, and they are explicitly provided, for example, by Drenth (1999).

6.8 Difference Fourier Technique

Supposing that one has solved the crystal structure of a biological macromolecule, and has isomorphous crystals of this macromolecule that contain small structural changes caused by substrate–analog or inhibitor binding, metal removal or replacement, or a local mutation of one or several amino acids, then these structural changes can be determined by the difference Fourier technique. The difference Fourier map is calculated by using the differences between the observed structure factor amplitudes of the slightly altered molecule $F_{DERI}(obs)$ and the native molecule $F_{NATI}(obs)$ as Fourier coefficients and the phase angles of the native molecule a_{NATI} as phases according to Eq. (6.30):

$$\rho_{DERI} - \rho_{NATI} \cong \frac{1}{V} \sum_{\mathbf{h}} m[F_{DERI}(obs) - F_{NATI}(obs)] \times \exp(i a_{NATI}) \exp(-2\pi i \mathbf{hx}) \tag{6.30}$$

where m may be the figure of merit or another weighting scheme. The difference Fourier map can alternatively be calculated with coefficients $F_{DERI}(obs)$, $F_{DERI}(calc)$ and phases $a_{DERI}(calc)$.

$F_{DERI}(calc)$ and $a_{DERI}(calc)$ do not include the unknown contribution of the structural change.

Figure 6.5 a and b illustrate the relationship for the structure factors involved in the difference Fourier technique. It is assumed that the structural change is small. If F_{NATI} is large, the structure factor amplitude of the structural change

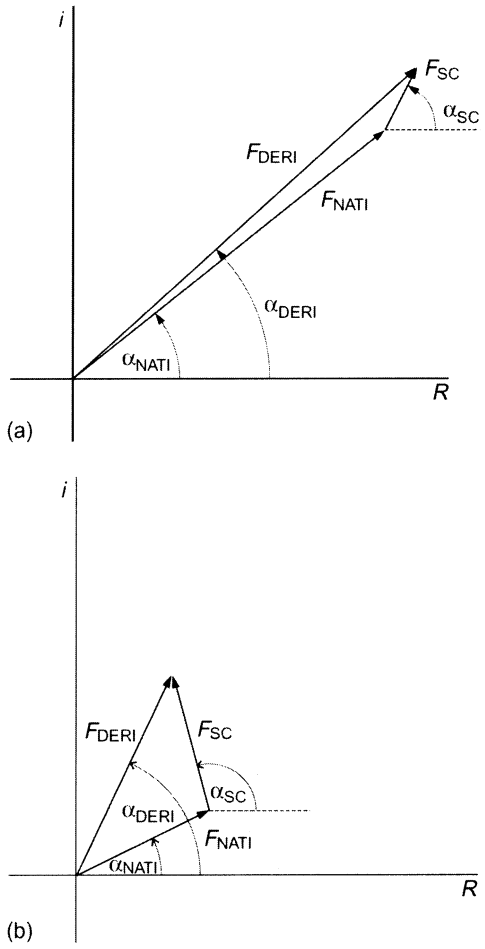


Fig. 6.5 Vector diagrams illustrating different situations (a) and (b) in the difference Fourier technique for the involved structure factors.

F_{SC} will be small compared with F_{NATI}, and a_{DERI} will be close to a_{NATI}. This is no longer valid if F_{NATI} is small. Now F_{SC} is comparable to F_{NATI}, and a_{DERI} may deviate considerably from a_{NATI}. This implies the necessity to introduce a weighting scheme that scales down the contributions where the probability is high that a_{NATI} differs appreciable from the correct phase angle. A similar situation is met if only a partial structure is known, for example from molecular replacement or preliminary model building. The structure factor $\mathbf{F}_P$ of the partial structures corresponds to the native structure factor $\mathbf{F}_{NATI}$, the structure factor $\mathbf{F}_N$ of the complete structure to that of the derivative structure factor $\mathbf{F}_{DERI}$, and the structure factor of the structural change $\mathbf{F}_{SC}$ to that of the structure factor $\mathbf{F}_Q$ of the missing atoms, all with the relevant amplitudes and phases. Various weighting schemes have been elaborated such as those of Sim (1959) and Read (1986). The weighting scheme of Sim has the following form (Eq. 6.31):

$$w = \frac{I_1(X)}{I_0(X)} \tag{6.31}$$

for acentric reflections and (Eq. 6.32):

$$w = \tanh\frac{X}{2} \tag{6.32}$$

for centric reflections with (Eq. 6.33)

$$X = \frac{2F_N \times F_P}{\varepsilon \sum_Q} \tag{6.33}$$

$I_0(X)$ and $I_1(X)$ are modified Bessel functions of zero*th* and first order, respectively. F_N is the structure amplitude of the entire structure and F_P that of the partial structure. The factor ε corrects for the difference in expected intensity for different zones in reciprocal space. The parameter $\sum_Q$ measures the amount of missing scattering matter, and is given by:

$$\sum_Q = \left\langle F_Q^2/\varepsilon \right\rangle . \tag{6.34}$$

We need an estimate of $\sum_Q$ from the known structure factor amplitudes F_N and F_P. Previously, Bricogne (1976) suggested $|\overline{F_N^2 - F_P^2}|$, while Read (1986) proposed $\overline{n(F_N - F_P)^2/\varepsilon}$ with $n = 1$ for centric and $n = 2$ for acentric reflections. The Sim weighting scheme has been derived for a partial structure with no errors. This is of course not true, and Srinivasan (1966) has developed a corresponding formula for a partial structure with errors. Hence, we obtain for X:

$$X = 2\sigma_A \frac{F_N}{\left(\sum_{j=1}^{N} f_j^2\right)^{1/2}} \frac{F_P}{\left(\sum_{i=1}^{P} f_i^2\right)^{1/2}} \Bigg/ (1 - \sigma_A^2) . \tag{6.35}$$

σ_A can be estimated by Eq. (6.36), as provided by Srinivasan (1966)

$$\sigma_A = \frac{\sum_{i=1}^{P} f_i^2}{\sum_{j=1}^{N} f_j^2} \langle \cos 2\pi \mathbf{h}\Delta\mathbf{r}_j \rangle_P \tag{6.36}$$

where $\Delta\mathbf{r}_j$ represents the finite errors in the coordinates of the calculated partial structure. Other approaches to estimate σ_A are discussed by Read (1986). Furthermore, a σ_A plot can be used to estimate the coordinate error of the structural model, which will be discussed in Section 7.3.2.3.2.

These equations and weighting schemes can also be used for the calculation of OMIT maps (where parts of the model have been omitted from the structure factor evaluation), or when a complete structure must be developed from a known partial model.

References

Abrahams, J.P., Leslie, A.G.W., *Acta Crystallogr.* **1996**, *D52*, 30–42.

Agarwal, R.C., Isaacs, N.W., **1977**, *74*, 2835–2839.

Bricogne, G., *Acta Crystallogr.* **1974**, *A30*, 395–405.

Bricogne, G., *Acta Crystallogr.* **1976**, *A32*, 832–847.

CCP4, *Acta Crystallogr.* **1994**, *D50*, 760–763.

Cowtan, K.D., Main, P., *Acta Crystallogr.* **1998**, *D54*, 487–493.

Cowtan, K.D., Main, P., *Acta Crystallogr.* **1993**, *D49*, 148–157.

Drenth, J., *Principles of Protein X-ray Crystallography*, Springer, New York, **1999**.

Gaykema, W.P.J., Hol, W.G.J., Verejken, J.M., Soeter, N.M., Bak, H.J., Beintema, J.J., *Nature* **1984**, *309*, 23–29.

Gaykema, W.P.J., Volbeda, A., Hol, W.G.J., *J. Mol. Biol.* **1985**, *187*, 255–275.

Hendrickson, W.A., Lattman, E.E., *Acta Crystallogr.* **1970**, *B26*, 136–143.

Kleywegt, G.J., Jones, T.A., Halloween … Masks and Bones. In: Bailey, S., Hubbard, R., Waller, R. (Eds.), *From First Map to final Model*. Daresbury Laboratory, Warrington, **1994**.

Ladenstein, R., Schneider, M., Huber, R., Bartunik, H., Schott, K., Bacher, A., *J. Mol. Biol.* **1988**, *203*, 1045–1070.

Lamzin, V.S., Wilson, K.S., *Methods Enzymol.* **1997**, *277*, 269–305.

Leslie, A.G.W., *Acta Crystallogr.* **1987**, *A43*, 134–136.

Löwe, J., Stock, D., Jap, B., Zwickl, P., Baumeister, W., Huber, R., *Science* **1995**, *268*, 533–539.

Main, P., *Acta Crystallogr.* **1990**, *A46*, 507–509.

Perrakis, A., Sixma, T.A., Wilson, K.S., Lamzin, V.S., *Acta Crystallogr.* **1997**, *D53*, 448–455.

Read, R.J., *Acta Crystallogr.* **1986**, *A42*, 140–149.

Rossmann, M.G., Arnold, E., Erikson, J.W., Frankenberger, E.A., Griffith, J.P., Hecht, H.-J., Johnson, J.E., Kamer, G., Luo, M., Mosser, A.G., Rueckert, R.R., Sherry, B., Vriend, G., *Nature* **1985**, *317*, 145–153.

Sayre, D., *Acta Crystallogr.* **1952**, *5*, 60–65.

Sim, G.A., *Acta Crystallogr.* **1959**, *12*, 813–815.

Srinivasan, R., *Acta Crystallogr.* **1966**, *20*, 143–144.

Vellieux, F.M.D., Hunt, J.F., Roy, S., Read, R.J., *J. Appl. Crystallogr.* **1995**, *28*, 347–351.

Wang, B.-C., *Methods Enzymol.* **1985**, *115*, 90–112.

Zhang, K.Y.J., *Acta Crystallogr.* **1993**, *D49*, 213–222.

Zhang, K.Y.J., Main, P., Histogram matching as a density modification technique for phase refinement and extension of protein molecules. In: Bailey, S., Dodson, E., Phillips, S. (Eds.), *Improving Protein Phases*, Report DL/SCI/R26, pp. 57–64. Daresbury Laboratory, Warrington, **1988**.

Zhang, K.Y.J., Main, P., *Acta Crystallogr.* **1990**, *A46*, 377–381.

Zhang, K.Y.J., Cowtan, K.D., Main, P. Phase improvement by iterative density modification. In: Rossmann, M.G., Arnold, E. (Eds.) *International Tables for Crystallography*, Vol. F, pp. 311–324. Kluwer Academic Publishers, Dordrecht, **2001**.

7
Model Building and Refinement

7.1
Model Building

Once the quality of the MIRAS or MAD maps is good enough, model building can be started. This is carried out using a computer graphics system, with the main modeling programs being "O" (Jones et al., 1991) and TURBO-FRODO (Jones, 1978; Roussel and Cambillau, 1989). An interesting alternative is the program MAIN (Turk, 1995), which additionally contains routines for molecular averaging, molecular docking, and other features. A new development is the program COOT (Emsley et al., 2004) which is now a supported program of the CCP4 program suite.

Visualization of the relevant electron density map on the computer graphics system appears as cage-like structures. For this purpose, the standard deviation from the mean value of the map is calculated and the cage-like structure is built up for a given contour level (normally 1.0 σ). The first task in a *de-novo* protein crystal structure analysis is to localize the trace of the polypeptide chain. This can be assisted by routines for automatic chain tracing such as BONES (Jones and Thirup, 1986), which is an auxiliary program of O. Such automatic chain-tracing programs generate a skeleton of the electron density map; this representation was introduced by Greer (1974). Automated chain-tracing modules are contained in the ARP/wARP program suite (Lamzin et al., 2001) and the program RESOLVE (Terwilliger, 2002). The method applied in ARP/wARP was briefly described in Section 6.5. RESOLVE constructs the model by template-matching and iterative fragment extension.

It should be noted that the ab initio model building is only successful if the resolution and phasing of the experimental electron density is sufficient (as a rule of thumb, better than 2 Å). Nonetheless, these techniques are very successful at lower resolutions when a partial model is available. Here, they use a hybrid model consisting of the partial and the free atom model.

Very often, the quality of the electron density map is inferior, such that the model cannot be constructed automatically. Consequently, the manual mode of model building has to be applied, and this is best started from a skeletonized version of the electron density map which is simultaneously displayed with the map. First, attempts are made to recognize secondary structure elements such

as α-helices or β-strands and β-sheets. When such pieces of the polypeptide chain have been identified, a Cα trace can be built into the electron density. A polyalanine chain (or segments of it) can be built from this trace, either by the use of databases or simply by assigning the identified secondary structure type. The atomic model is represented as sticks which connect the atomic centers of bonded atoms. The individual building blocks (amino acids) of the protein molecule can be generated, interactively manipulated (e.g., linked with each other, moved, rotated, etc.), and then fitted into the corresponding part of the electron density map. The geometry of the atomic model is regularized according to protein standard geometries. At this stage, the direction of the polypeptide chain, or of its segments, may be incorrectly assigned. Nevertheless, such partial models can be used to improve the structure factor phases by refining them crystallographically against the observed structure factor amplitudes. This phase information can be used directly to calculate a new electron density map, commonly with $2F_{obs}–F_{calc}$ Fourier coefficient amplitudes. This type of map is the sum of a normal F_{obs} Fourier and a difference Fourier synthesis. It displays the atomic model with normal weight, and also indicates errors in the model by its contribution of the difference Fourier map. The parallel determination and inspection of a difference Fourier map is also very helpful. As mentioned previously, the model phases can be combined with phases present from other sources or incorporated into procedures of phase improvement. A further model-building cycle can be started with such new and improved electron density maps. The quality of the maps should now improve in such a way that side chains can be correctly assigned; this allows the correct direction of the polypeptide chain or of its segments to be determined, and their position in the corresponding amino acid sequence to be located. After several cycles of model building and crystallographic refinement the atomic model will be so well defined that the solvent structure of internally bound solvent molecules can be developed. The atomic model is now complete and the biochemical interpretation can be started.

The Uppsala Software Factory (USF, http://xray.bmc.uu.se/usf/) has developed a suite of programs around "O", which are very useful in model building, refinement, and related topics. These programs have been listed and described briefly by their authors (Kleywegt et al., 2001). The first such program is RAVE, a suite of programs for electron density improvement and analysis, with a strong focus on averaging techniques (Kleywegt and Read, 1997). RAVE contains, for example, the following important programs:

- AVE is used for averaging and expanding the averaged electron density according to the crystal symmetry;
- COMA serves to calculate local density correlation maps that can be used to define masks (molecular envelopes);
- IMP optimizes NCS operators by relating two copies of a molecule (or domain) inside the same cell;
- NCS6D may be very helpful to find NCS operators in cases where it is difficult to obtain them by other means.

The program rotates and translates a skeletonized version (BONES atoms or atoms in PDB coordinate format) of the electron density map against the same map in all six dimensions, and then calculates the correlation coefficient between the transformed atoms and the density around the atoms. RAVE further contains tools for averaging between crystal forms:

- MASKIT calculates a local density correlation map from the density of the two different crystal forms using the algorithm of Read (Vellieux et al., 1995);
- MAVE performs the (skew) density averaging and expansion steps, but now separately because the density of the various crystal forms has also to be averaged. This program also contains the option to improve operators that relate the position and orientation of the molecular envelope (mask) in one crystal form with those in the other crystal form.
- CONDEM combines the individual (possibly averaged) density from various crystal forms.

Finally, two utility programs of RAVE should be mentioned here. The first is MAMA, a program used to generate, analyze, and manipulate masks. Masks can, for example, be generated from scratch using BONES atoms, a structural model contained in a PDB-file or defining a sphere or box around a given point in the map. The second program, MAPMAN, is used for format conversion, analysis and manipulation of electron density maps. Maps can be read and written in a variety of formats including those used by "O".

The USF supplies many other utility programs for macromolecular structure analysis, and two of these will be mentioned here in more detail.

- LSQMAN (Kleywegt, 1996; Kleywegt and Jones, 1997) is a program used for analyzing and manipulating copies of a molecule or multiple molecules. It contains tools to superimpose molecules (including an option to find such superpositioning automatically, which works very efficiently), to improve the fit of both superimposed molecules in many different ways, to calculate and plot r.m.s. distances, and several other tools. The program can handle proteins, nucleic acids, and other types of molecules.
- MOLEMAN2 is a general program used for the analysis and manipulation of molecules (in PDB-format files). Although this program contains too many options to list here, the tools to analyze and manipulate temperature factors and occupancies, to orthogonalize, deorthogonalize or shift atomic coordinates or to renumber the residues in the coordinate set should be especially noted, though all of the other options are also very beneficial.

Model building and refinement are closely linked to each other. Methods of crystallographic refinement are covered in the following section.

7.2 Crystallographic Refinement

7.2.1 Introduction

The structural model has to be subjected to a refinement procedure. Macromolecular crystallography does not differ fundamentally from small-molecule crystallography, but is complicated by several peculiarities. First, typical macromolecules contain thousands of atoms and crystallize in unit cells with cell lengths of up to several hundred Angströms. This causes a high number of intensity data in an X-ray experiment, and a large set of parameters to be refined. Moreover, whilst small-molecule refinement programs were simply not designed for such large structures, the capability of computers was also inadequate some 30 years ago. Second, the resolution of the diffraction data is normally below atomic resolution, which does not allow any application of the least-squares technique for the parameter refinement – the standard technique used for small molecules. Therefore, the single atoms cannot be treated as moving independently; rather, they must be refined using energy or stereochemistry restraints, taking care to maintain a reasonable stereochemistry of the macromolecule. This reduces the number of parameters to be refined considerably and places a reasonable value on the ratio of observations to parameters to be refined. Nowadays, computer performance is high enough also to refine macromolecular structures at atomic resolution. Indeed, this has been reflected by an extension of the crystallographic refinement program SHELXL (Sheldrick and Schneider, 1997) to treat macromolecules.

"Improving the agreement" between the observed and calculated data can be done by different criteria to measure the agreement. The most commonly used measure is the L_2 norm of the residuals, which is simply the sum of the squares of the differences between observed and calculated data,

$$L_2(\mathbf{x}) = ||w_i[y_i - M_i(\mathbf{x})]|| = \sum_i w_i[y_i - M_i(\mathbf{x})]^2 \,, \tag{7.1}$$

where w_i is the weight of observation, y_i and $M_i(\mathbf{x})$ is the calculated value of observation i given the parameters $\mathbf{x}$. The least-squares refinement now searches for the set of model parameters that give the minimum variance of the observations. One problem of the L_2 norm is that it is very sensitive against large deviations, which occur especially during the early stages of refinement. To overcome this, it may be better to refine against the L_1 norm:

$$L_1(\mathbf{x}) = \sum_i w_i|[y_i - M_i(\mathbf{x})]| \,, \tag{7.2}$$

which is the sum of the absolute value of the residuals. However, to date this technique has not been used in macromolecular crystallography.

Many of the problems of least-squares refinement can be overcome by changing the measure of agreement from least squares to maximum likelihood. The model is adjusted to maximize the probability of the given observations. Maximum likelihood refinement is particularly useful for incomplete models because it produces residuals that are less biased by the current model than those obtained by least squares. Maximum likelihood also provides a rigorous formulation for all forms of error in both the model and the observations, and allows incorporation of additional forms of prior knowledge (such as additional phase information) into the probability distributions.

The likelihood of a model represented by a set of observations is the product of the probabilities of all the observations of the given model. As in our case, the observations are structure factors $\mathbf{F}_i$ with a conditional probability distribution $P(\mathbf{F}_i; \mathbf{F}_{i,c})$, the likelihood of the model becomes

$$L = \prod_i P(\mathbf{F}_i; \mathbf{F}_{i,c}) \ , \tag{7.3}$$

where $\mathbf{F}_{i,c}$ is the calculated model structure factor. This is usually transformed in its logarithmic form:

$$\log L = \sum_i \log P(\mathbf{F}_i; \mathbf{F}_{i,c}) \ , \tag{7.4}$$

which is more tractable.

7.2.2 Principles of Least Squares

A detailed description of the principles of the least-squares technique has been given by Prince and Boggs (1999). Here, we will present an outline of these studies to provide a basic understanding of the method. The method of least squares may be formulated as follows: Given a set of n observations, $y_i \, (i = 1, 2, \ldots, n)$, that are measurements of quantities that can be described by differentiable model functions, $M_i(\mathbf{x})$, where $\mathbf{x}$ is a vector of parameters, $x_i (i = 1, 2, \ldots, p)$ find the values of the parameters for which the sum L_2 (Eq. 7.1) is minimum. The values of the parameters that deliver the minimum of L_2 are called *estimates* of the parameters, and a function of the data that locates the minimum is an *estimator*. The necessary condition for L_2 to be a minimum is for the first derivation to vanish, which results in a set of simultaneous equations, the *normal equations*, of the form:

$$\partial L_2 / \partial x_j = -2 \sum_{i=1}^{n} [y_i - M_i(\mathbf{x})] \partial M_i(\mathbf{x}) / \partial x_j = 0 \ . \tag{7.5}$$

The model functions, $M_i(\mathbf{x})$, are, in general, nonlinear, and there are no direct ways to solve these systems of equations. Iterative methods must be used to

solve them. In many cases, linear approximations to the model functions are good approximations in the vicinity of the minimum. The linear approximation of the ith model function can now be written as:

$$M_i(\mathbf{x}) \approx b_i + \sum_{j=1}^{p} A_{ij} x_j \,, \tag{7.6}$$

where A_{ij} are the elements of a matrix $\mathbf{A}$ and b_i are the elements of a vector $\mathbf{b}$. We can write Eq. (7.6) in matrix form with column vectors $\mathbf{M}(\mathbf{x})$ (Eq. 7.7) and $\mathbf{y}$, whose ith elements are $M_i(\mathbf{x})$ and y_i,

$$\mathbf{M}(\mathbf{x}) \approx \mathbf{b} + \mathbf{A}\mathbf{x} \,, \tag{7.7}$$

and for this linear model, L_2 (Eq. (7.8) becomes

$$L_2 = [(\mathbf{y} - \mathbf{b}) - \mathbf{A}\mathbf{x}]^{\mathrm{T}} \mathbf{W} [(\mathbf{y} - \mathbf{b}) - \mathbf{A}\mathbf{x}] \,, \tag{7.8}$$

where $\mathbf{W}$ is a diagonal matrix whose diagonal elements are $W_{ii} = w_i$. In this notation the normal Eq. (7.5) can be written as

$$\mathbf{A}^{\mathrm{T}}\mathbf{W}\mathbf{A}\mathbf{x} = \mathbf{A}^{\mathrm{T}}\mathbf{W}(\mathbf{y} - \mathbf{b}) \,, \tag{7.9}$$

and their solution is

$$\hat{\mathbf{x}} = (\mathbf{A}^{\mathrm{T}}\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{W}(\mathbf{y} - \mathbf{b}) \,. \tag{7.10}$$

If $W_{ii} > 0$ for all I and $\mathbf{A}$ has full column rank, then $\mathbf{A}^{\mathrm{T}}\mathbf{W}\mathbf{A}$ will be positive definite, and L_2 will have a unique minimum at $\mathbf{x} = \hat{\mathbf{x}}$. The matrix $\mathbf{H} = (\mathbf{A}^{\mathrm{T}}\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{W}$ is a $p \times n$ matrix that relates the n-dimensional observation space to the p-dimensional parameter space and is known as *least-squares estimator*. The least-squares estimator has some special properties in statistical analysis, but the reader is referred to Prince and Boggs (1999) for a treatment of this issue.

The general case is a nonlinear relation between the model parameters and the observations, and this holds for crystallographic refinement. We can recall that L_2 (Eq. (7.1) was valid for the general nonlinear case, but now $M_i(\mathbf{x})$ is given by a nonlinear relation. In Prince and Boggs (1999), two useful ways of solving the minimization problem are described, the Gauss-Newton algorithm and the Quasi-Newton method. We will briefly explain the first method. Let $\mathbf{x}_c$ be the current approximation $\hat{\mathbf{x}}$, the solution of Eq. (7.1). We construct a linear approximation to $M_i(\mathbf{x})$ in the vicinity of $\mathbf{x} = \mathbf{x}_c$ by expanding in a Taylor's series through the linear terms, obtaining

$$M_i(\mathbf{x}) \approx M_i(\mathbf{x}_c) + \sum_{j=1}^{p} J_{ij}(\mathbf{x} - \mathbf{x}_c)_j \,, \tag{7.11}$$

where **J** is the Jacobian matrix, defined by

$$J_{ij} = \partial M_i(\mathbf{x})/\partial x_j \ . \tag{7.12}$$

A straightforward procedure, known as the *Gauss-Newton algorithm*, may be formally stated as follows:
1. Compute **d** as the solution of the linear system:

$$\mathbf{J}^{\mathrm{T}}\mathbf{W}\mathbf{J}\mathbf{d} = \mathbf{J}^{\mathrm{T}}\mathbf{W}[\mathbf{y} - \mathbf{M}(\mathbf{x}_c)] \ . \tag{7.13}$$

2. Set $\mathbf{x}_c = \mathbf{x}_c + \mathbf{d}$.
3. If not converged, go to (1), else stop.

The convergence rate of the Gauss-Newton algorithm (Dennis and Schnabel, 1983) depends on the size of the residual, that is, on $L_2(\hat{\mathbf{x}})$. If $L_2(\hat{\mathbf{x}}) = 0$, then the convergence rate is quadratic, but if it is small then the rate is linear; however, if $L_2(\hat{\mathbf{x}})$ is large, then the procedure is not locally convergent at all. Fortunately, this procedure can be changed so that it is always locally convergent, and even globally convergent – that is, convergent to a relative minimum from a starting point (for details see Prince and Boggs, 1999). The basis for an understanding of the least-squares method has been laid down and we can now pass over to the crystallographic refinement itself.

7.2.3
Constraints and Restraints in Refinement

The techniques of least squares are applicable for refining almost any model, but the question of suitability of the model remains. In many cases, the model implies constraints, the application of which constricts the solutions. The classical technique for the application of constraints is to use Lagrange undetermined multipliers, but in this approach the set of p parameters x_j is increased by $p - q (q < p)$ additional unknowns λ_k, one for each constraint relationship desired. This is not a beneficial situation for macromolecular refinement, where a reduction of parameters is desirable by the application of restraints. In most cases encountered in crystallography the constraints may be applied directly, thus reducing rather than increasing the size of the normal equations matrix. For each constraint introduced one of the parameters becomes dependent on the remaining set, and the rank of the remaining system is reduced by one. For p parameters and $p - q$ constraints, the problem reduces to q parameters. Using the Gauss-Newton algorithm, the normal-equations matrix is $\mathbf{A}^{\mathrm{T}}\mathbf{W}\mathbf{A}$, where

$$A_{ij} = \partial M_i/\partial x_j \ , \tag{7.14}$$

and **W** is a weight matrix. The constraint relations may be written as

$$x_k = f_k(z_1, z_2, \ldots, z_q) \ , \tag{7.15}$$

where the zs are the parameters of the constrained model. Applying the chain rule for differentiation, the normal-equations matrix for the constrained model becomes $\mathbf{B}^{\mathrm{T}}\mathbf{W}\mathbf{B}$, where

$$B_{ik} = \partial M_i(\mathbf{x})/\partial z_k = \sum_{j=1}^{p} [\partial M_i(\mathbf{x})/\partial x_j](\partial x_j/\partial z_k) \ . \tag{7.16}$$

This can be written in matrix form $\mathbf{B} = \mathbf{AC}$, where $C_{ij} = \partial x_j/\partial z_k$ defines a $p \times q$ constraint matrix. The application of constraints involves: (i) determination of the model to be used; (ii) calculation of the elements of **C**; and (iii) computation of the modified normal-equations matrix.

Most existing programs calculate the structure factor amplitude F and its partial derivatives with respect to the positional parameters, isotropic or anisotropic temperature factors, occupancy and an anisotropic overall scale factor. The constrained calculation is usually made by evaluating selected elements, $\partial x_j/\partial z_k$. The following constraints must taken into account in such crystallographic refinement programs:

- the crystal structure contains atoms in special positions or positional parameters are linearly dependent on others, which may occur in trigonal, hexagonal, tetragonal and cubic space groups;
- occupancies of certain sites in the crystal; and
- some portion of the structure undergoes thermal motion as a rigid.

The reader is referred to Prince et al. (1999) for more details in the mathematical treatment of these special constraints.

We already mentioned the necessary degree of overdetermination to obtain an accurate structural model. For well-ordered crystals of small- and intermediate-sized molecules it is usually possible to measure a hundred or more independent reflections for each symmetry-independent atom. With ten parameters per atom (three positional, six anisotropic B-factors, one occupancy factor) the overdeterminacy is still greater than 10 to 1. The situation is different with X-ray studies of biological macromolecules, where the number of independent reflections is often fewer than the number of parameters necessary to define the distributions of individual atoms. This problem may be overcome either by reducing the number of parameters describing the model, or by increasing the number of independent observations. Both approaches utilize knowledge of stereochemistry. Above, we have discussed the use of constraints to introduce this stereochemical knowledge. Now, we explain a technique that introduces the stereochemical conditions as additional observational equations. This method differs from the other in that information is introduced in the form of distributions about mean values rather than as rigidly fixed geometries. The parameters are restrained to fall within energetically permissible limits.

For restrained refinement the sum L_2 (Eq. (7.1) to be minimized now contains several classes of observational equations, in addition to those for structure

factors. The stereochemical restraints are usually introduced as energy terms (e.g., Jack and Levitt, 1978) or by expressing all types of stereochemical restraints as distances (e.g., Ten Eyck et al., 1976; Konnert and Hendrickson, 1980).

The minimization of a potential energy function E together with a diffraction term D is done according to:

$$L_2 = E + D \tag{7.17}$$

where

$$E = \sum k_b[b_{j(\text{calc})} - b_j^0]^2 + \sum k_\tau[\tau_{j(\text{calc})} - \tau_j^0]^2 + \sum k_\Theta[1 + \cos(m\Theta_k + \delta)] + \sum (Ar^{-12} - Br^{-6}) \tag{7.18}$$

$$D = \sum_i w_i[F_{i(\text{obs})} - kF_{i(\text{calc})}]^2 \tag{7.19}$$

and is applied in the programs EREF (Jack and Levitt, 1978) and CNS (Brünger et al., 1998), which is now used frequently. The four terms of the right-hand side of E describe bond, valence angle, dihedral torsion angle, and non-bonded interactions, k_b is the bond stretching constant, k_τ is the bond angle bending force constant, k_Θ is the torsional barrier, m and δ are the periodicity and phase of the barrier, A and B are the repulsive and long-range nonbonded parameters, D is the crystallographic contribution with w_i a weighting factor, F_{obs} the observed structure factor, F_{calc} the calculated structure factor, and k a scaling factor.

Stereochemical restraints as distances are used in the programs PROLSQ (Hendrickson, 1985) and TNT (Tronrud et al., 1987). The sum to be minimized adopts the following form:

$$L_2 = \sum_i w_i[F_{i(\text{obs})} - kF_{i(\text{calc})}]^2 \quad \text{(i)} \tag{7.20}$$

$$+ \sum_{\text{dist},j} w_{\text{D},j}(d_j^{\text{ideal}} - d_j^{\text{model}})^2 \quad \text{(ii)}$$

$$+ \sum_{\substack{\text{planes} \\ k}} \sum_{\substack{\text{coplanar} \\ \text{atoms } i}} w_{\text{P},i,k}(\mathbf{m}_k \cdot \mathbf{r}_{i,k} - d_k)^2 \quad \text{(iii)}$$

$$+ \sum_{\substack{\text{chiral} \\ \text{centers } l}} w_{\text{C},l}(V_{\text{C},l}^{\text{ideal}} - V_{\text{C},l}^{\text{model}})^2 \quad \text{(iv)}$$

$$+ \sum_{\substack{\text{nonbonded} \\ \text{contacts } m}} w_{\text{N},m}(d_m^{\min} - d_m^{\text{model}})^4 \quad \text{(v)}$$

$$+ \sum_{\substack{\text{torsion} \\ \text{angles } t}} w_{\Gamma,t}(X_t^{\text{ideal}} - X_t^{\text{model}})^2 \qquad \text{(vi)}$$

where (i) is the conventional structure factor term. The distances d_j between bonded atoms and between next-nearest-neighbor atoms are used to require bonded distances and angles to fall within acceptable ranges (term (ii)). Groups of atoms may be restrained to be near a common plane (Schomaker et al., 1959) (term (iii)), where $\mathbf{m}_k$ is the unit vector normal to the plane, $\mathbf{r}_{i,k}$ is the position of an atom and d_k is the distance of the least-squares plane from the origin. $\mathbf{m}_k \cdot \mathbf{r}_{i,k} - d_k$ is then the distance of the atom from the least-squares plane.

Interatomic distances are independent of the handedness of enantiomorphous groups such as C_α-groups (except of glycine) or C_β of threonine and isoleucine in proteins. If $\mathbf{r}_c$ is the position vector of a central atom and $\mathbf{r}_1$, $\mathbf{r}_2$ and $\mathbf{r}_3$ are the positions of three atoms bonded to it, such that the four atoms are not coplanar, the chiral volume is defined by

$$V_C = (\mathbf{r}_1 - \mathbf{r}_c) \cdot [(\mathbf{r}_2 - \mathbf{r}_c) \times (\mathbf{r}_3 - \mathbf{r}_c)] \,, \qquad (7.21)$$

where $\times$ indicates the vector product. The chiral volume may be either positive or negative, depending on the handedness of the group. The expression to be refined for the chiral volumes is given in term (iv). Contacts between nonbonded atoms are important for determining the conformation of folded, chain molecules and crystal packing. They have been modeled by a potential function that is strongly repulsive when the interatomic distance is less than some minimum value, but only weakly attractive, so that it can be neglected in practice, when the distance is greater than this value. This leads to an expression of the form of term (v), which is included only if $d_m^{\text{min}} < d_m^{\text{model}}$. Macromolecules usually retain flexibility by relatively unrestricted rotations about single bonds. There are, nevertheless, significant restrictions of these torsion angles X_t, which can be restrained by terms in the form of (vi). The torsion angles are dihedral angles between planar groups at opposite ends of the bond. The weighting factors in Eq. (7.29) are usually chosen as $w = 1/\sigma^2$, except for w_N, which adopts the form of $1/\sigma^4$, where σ is the standard deviation of the expected distribution. A specific feature of the program TNT (Tronrud et al., 1987) is the calculation of the gradients of the model function via fast Fourier transforms (FFT), which makes it very efficient.

For both refinement schemes, parameters are employed which were derived from small-molecule crystal structures of amino acids, small peptides, nucleic acids, sugars, fatty acids, cofactors, etc. (Engh and Huber, 1991). If NCS symmetry is present, a corresponding term may be introduced into the energy or stereochemistry part of the expression to be minimized. It is possible to divide the structural model into several individual parts and to refine these parts as rigid bodies. This is especially useful with solutions from molecular replacement.

A measure of the quality of the crystallographic model is calculated from the crystallographic *R*-factor:

$$R = \frac{\sum_i |F_{\text{obs}}| - |k|F_{\text{calc}}|}{\sum_i |F_{\text{obs}}|} \tag{7.22}$$

Typical *R*-factors are below 0.2 for a well-refined macromolecular structure.

Beside the atomic coordinates *x*, *y*, *z*, the atomic temperature factor *B* may be refined at a resolution better than 3.5 Å. This is done in most programs in a separate step where, for example, in program CNS the target function

$$T = E_{\text{XRAY}} + E_R \tag{7.23}$$

is minimized, where

$$E_R = W_B \sum_{(i,j)\text{-bonds}} \frac{(B_i - B_j)^2}{\sigma^2_{\text{bonds}}} + W_B \sum_{(i,j,k)\text{-angles}} \frac{(B_i - B_k)^2}{\sigma^2_{\text{angles}}} + W_B \times \sum_{k\text{-group}} \sum_{j\text{-equivalences}} \sum_{i\text{-unique atoms}} \frac{(B_{ijk} - \overline{B_{ijk}})^2}{\sigma^2_{\text{ncs}}} \tag{7.24}$$

The last term is used only if NCS symmetry restraints should be imposed on the molecules. Normally, isotropic *B*-factors are applied and refined in macromolecular crystallography only. Even for a high-resolution structure (1.7 Å), the ratio of observations (observed structure factors) to parameters to be refined (*x*, *y*, *z*, *B* for each atom) is only about 3. Therefore, as already mentioned, as many as possible additional "observations" (energy or stereochemistry restraints) are incorporated. In some cases it is useful to refine the individual occupancy of certain atoms such as bound metal ions or solvent atoms. This must be performed in a separate step.

7.2.4
Refinement by Simulated Annealing

All of the above-mentioned refinement procedures are based on the least-squares method. The radius of convergence for this method is not very high because it follows a downhill path to its minimum. If the model is too far away from the correct solution, the minimization may end in a local minimum corresponding to an incorrect structure. Brünger and colleagues (Brünger et al., 1987, 1993) introduced the method of simulated annealing (SA), which is able to overcome barriers in the L_2-function and find the correct global minimum. The SA algorithm requires a mechanism to create a Boltzmann distribution at a given temperature T and an annealing schedule $T_1 \geq T_2 \geq \ldots \geq T_l$ at which the Boltzmann distribution is computed. There exist several methods to be

used. For crystallographic refinement, molecular dynamics (MD) has proven extremely successful (Brünger et al., 1987) because it limits the search to reasonable "moves". A suitably chosen set of atomic parameters can be considered as generalized coordinates that are propagated in time by the classical equations of motion (Goldstein, 1980). If the generalized coordinates represent the x, y, z positions of the atoms of a molecule, the classical equations of motion reduce to the well-known Newton's second law:

$$m_i \frac{\partial^2 \mathbf{r}_i}{\partial t^2} = -\nabla_i E \,, \tag{7.25}$$

where the quantities m_i and $\mathbf{r}_i$, respectively, are the mass and coordinates of atom i, and E is a target function as given by Eq. (7.17). The crystallographic term D is treated as a pseudoenergy term. The solution of the partial differential Eq. (7.25) can be achieved by finite difference methods; this approach is denoted as molecular dynamics.

The initial velocities for the integration of Eq. (7.25) are usually assigned randomly from a Maxwell distribution at the appropriate temperature. Assignment of different initial velocities will generally produce a somewhat different structure after SA. By performing several different initial velocities, one can therefore improve the chances of success of SA refinement. The Cartesian MD applies chemical restraints contained in the energy term E of Eq. (7.17). One can transform these restraints into constraints (e.g., fixed bond lengths and bond angles) by using torsion angle MD, which was introduced for crystallographic refinement by Rice and Brünger (1994). This method is numerically very robust and has a significantly increased radius of convergence in crystallographic refinement compared to Cartesian MD (Rice and Brünger, 1994).

SA requires the control of the temperature during the MD. The current temperature of the simulation (T_{curr}) is calculated from the kinetic energy:

$$E_{\text{kin}} = \sum_i^n \frac{1}{2} m_i \left(\frac{\partial r_i}{\partial t}\right)^2 \tag{7.26}$$

of the MD simulation,

$$T_{\text{curr}} = 2E_{\text{kin}}/3nk_B \,, \tag{7.27}$$

where n is the number of atoms and k_B is Boltzmann's constant. One commonly used approach to control the temperature of the simulation consists of coupling the equation of motions to a heat bath through a "friction" term (Berendsen et al., 1984). Another approach is to rescale periodically the velocities in order to match T_{curr} with the target function.

SA refinement is capable, for example, of overcoming a high-energy barrier occurring in the flipping of a peptide plane. It can be useful in removing model bias from the system. Multi-start refinement and structure factor averaging may give improved results (Rice et al., 1998).

7.2.5
The Maximum Likelihood Method

In Section 7.2.1, the maximum likelihood method was briefly introduced as another possibility for the crystallographic refinement of macromolecular structures. The basic assumption in least-squares minimization is that the conditional distribution of each F_{obs} or I_{obs} when the model is known as Gaussian with the expected value F_{obs} or I_{obs} and known uncertainties. In Eq. (7.3), the conditional probability distribution was used for the observed structure factor $\mathbf{F}_{obs}$. In an X-ray diffraction experiment, we determine the amplitude F_{obs} of the structure factor only. Since likelihood is proportional to the conditional probability distribution of experimental data when the model is known, the form of this conditional probability is needed. The best way to do this would be to find the joint probability distribution of all structure factors, but this task is not trivial and requires a large amount of computer memory and time. Therefore, all existing refinement procedures assume that the errors in different reflections are independent, and this simplification still delivers useful results.

With this assumption, the required joint probability distribution of all structure factor amplitudes has the form

$$P\left[(F_{obs})^{\text{all reflections}}; (\mathbf{F}_{calc})^{\text{all reflections}}\right] = \prod_{\text{all reflections}} P(F_{obs,i}; \mathbf{F}_{calc,i}) \ . \tag{7.28}$$

Thus, to describe the likelihood function, the conditional probability distribution of each reflection is generated and these are multiplied together to give the joint probability distribution.

As mentioned above, the ratio of the number of experimental data to the number of parameters to be estimated is low, and therefore prior stereochemical or other information must be used. This means that, from a mathematical point of view, all macromolecular refinement can be seen as the application of Bayes' theorem.

With experimental data F_{obs} and parameters to be estimated x, Bayes' theorem can be written as

$$P(\mathbf{x}; F_{obs}) = p(\mathbf{x})P(F_{obs}; \mathbf{x})/P(F_{obs}) = p(\mathbf{x})L(\mathbf{x}; F_{obs}) \ . \tag{7.29}$$

Here, P is the posterior probability distribution of the parameters when the experimental data are known; $\mathbf{x}$ are the parameters to be estimated; and p is the prior probability distribution of the parameters known before the experiment. L is the likelihood function which is proportional to the conditional distribution of experimental data when the parameters are known.

To best estimate the parameters $\mathbf{x}$, the posterior probability distribution function must be forced to reach its maximum. In order to apply this theorem, the form of the prior probability distribution and the likelihood is needed. $P(F_{obs}; \mathbf{x})$ and $L(\mathbf{x}; F_{obs})$ will be applied as $P(F_{obs}; \mathbf{F}_{calc})$ and $L(\mathbf{F}_{calc}; F_{obs})$, respectively, as $\mathbf{F}_{calc}$ is directly calculated from the $\mathbf{x}$.

Since maximization of a function is equivalent to the minimization of its negative logarithm, we obtain from Eqs. (7.29) and (7.28) the log-likelihood function LLK:

$$\mathrm{LLK} = -\log P(\mathbf{x}; (F_{\mathrm{obs}})^{\text{all reflections}}) = -\log p(\mathbf{x}) - \sum_{\text{all reflections}} \log L(\mathbf{F}_{\mathrm{calc},i}; F_{\mathrm{obs},i}) \ , \tag{7.30}$$

where $L(\mathbf{F}_{\mathrm{calc},i}; F_{\mathrm{obs},i}) \propto P(F_{\mathrm{obs},i}; \mathbf{F}_{\mathrm{calc},i})$. $-\log p(\mathbf{x})$ can be written in a straightforward way from stereochemical information; here, only the second term has to be taken into account. The minimization of the function on the left-hand side of Eq. (7.30) represents the amplitude-based maximum-likelihood (MLKF) residual.

We will not describe the detailed derivation of the actual LLK function used in programs BUSTER (Roversi et al., 2000), REFMAC (Murshudov et al., 1997) and CNS (Brünger et al., 1998). We state here the conditional probability distribution for the structure factor amplitude and the respective log likelihood function as given in Murshudov et al. (1997):

$$P(F_{\mathrm{obs}}; (\mathbf{F}_{\mathrm{calc},j})^{j=1,N_{\mathrm{part}}}) = \begin{cases} \dfrac{2F_{\mathrm{obs}}}{2\sigma^2_{F_{\mathrm{obs}}} + \Sigma_{\mathrm{WC}}} \exp - \left(\dfrac{F^2_{\mathrm{obs}} + F^2_{\mathrm{WC}}}{2\sigma^2_{F_{\mathrm{obs}}} + \Sigma_{\mathrm{WC}}} \right) I_0 \left(\dfrac{2F_{\mathrm{obs}} F_{\mathrm{WC}}}{2\sigma^2_{F_{\mathrm{obs}}} + \Sigma_{\mathrm{WC}}} \right) & \text{acentric} \\ \left[\dfrac{2}{\pi(\sigma^2_{F_{\mathrm{obs}}} + \Sigma_{\mathrm{WC}})} \right] \exp - \left[\dfrac{F^2_{\mathrm{obs}} + F^2_{\mathrm{WC}}}{2(\sigma^2_{F_{\mathrm{obs}}} + \Sigma_{\mathrm{WC}})} \right] \cosh \left(\dfrac{2F_{\mathrm{obs}} F_{\mathrm{WC}}}{2\sigma^2_{F_{\mathrm{obs}}} + \Sigma_{\mathrm{WC}}} \right) & \text{centric} \end{cases} . \tag{7.31}$$

where $\Sigma_{\mathrm{WC}} = \varepsilon \sum_{j=1}^{N_{\mathrm{part}}} \sum_j (1 - \mathbf{D}_j^2)$, $\mathbf{D}_j = \langle \exp[-\Delta B_j S^2/4] \cos 2\pi S \Delta x_j \rangle$, Δx_j is the error in position of atoms in the jth partial structure, ΔB_j is the error in B values of atoms in the jth partial structure, ε is the multiplicity of the diffraction plane, N_{part} is the number of partial structures, $\sigma_{F_{\mathrm{obs}}}$ is the experimental uncertainty in the structure factor amplitude, and F_{WC} is the amplitude of the weighted sum of partial calculated structure factors:

$$\mathbf{F}_{\mathrm{WC}} = \sum_{j=1}^{N_{\mathrm{part}}} \mathbf{D}_j \mathbf{F}_{\mathrm{calc},j}$$

$$\mathrm{LLK} = \sum_i \mathrm{LLK}_i \ , \tag{7.32}$$

where

$$\mathrm{LLK}_i = \begin{cases} c_a - \log F_{\mathrm{obs}} + \log(2\sigma^2_{F_{\mathrm{obs}}} + \Sigma_{\mathrm{WC}}) + \dfrac{F^2_{\mathrm{obs}} + F^2_{\mathrm{WC}}}{2\sigma^2_{F_{\mathrm{obs}}} + \Sigma_{\mathrm{WC}}} - \log I_0\left(\dfrac{2F_{\mathrm{obs}}F_{\mathrm{WC}}}{2\sigma^2_{F_{\mathrm{obs}}} + \Sigma_{\mathrm{WC}}}\right) & \text{acentric} \\ c_c + \frac{1}{2}\log(\sigma^2_{F_{\mathrm{obs}}} + \Sigma_{\mathrm{WC}}) + \dfrac{F^2_{\mathrm{obs}} + F^2_{\mathrm{WC}}}{2(\sigma^2_{F_{\mathrm{obs}}} + \Sigma_{\mathrm{WC}})} - \log\cosh\left(\dfrac{F_{\mathrm{obs}}F_{\mathrm{WC}}}{(\sigma^2_{F_{\mathrm{obs}}} + \Sigma_{\mathrm{WC}})}\right) & \text{centric} \end{cases} \tag{7.33}$$

and c_a and c_c are the respective expressions for stereochemical information for acentric and centric reflections.

If prior phase information is available it can be incorporated into the maximum likelihood expressions (Bricogne and Irwin, 1996; Murshudov et al., 1997; Pannu et al., 1998). This additional information strengthens the maximum likelihood structure refinement. In summary, the results derived using the maximum likelihood residual are consistently better than those from least-squares refinement.

7.2.6 Refinement at Atomic Resolution

If the resolution of a biological macromolecular crystal structure is equal to or better than 1.2 Å with at least 50% of the intensities in the outer shell being higher than 2σ, it is in the range of real atomic resolution and the ratio of observations to parameters is high enough to carry out, in principle, an unrestrained crystallographic refinement (Sheldrick, 1990). Each atom can be described by up to 11 parameters: the atom type, fixed after identification; three positional parameters (x, y, z); one isotropic atomic displacement parameter (ADP) or six anisotropic ADPs; one occupancy factor. At atomic resolution, six anisotropic displacement factors are used in the same way as in Eq. (3.81). The thermal-ellipsoid model is used to represent ADPs (Fig. 7.1). These reflect both the thermal vibrations of about the mean position as a function of time (dynamic disorder) and the variation of positions between different unit cells of the crystal arising from its imperfection (static disorder). Following Murshudov et al. (1999), the apparent ADP (U_{atom}) may be composed as:

$$U_{\mathrm{atom}} = U_{\mathrm{crystal}} + U_{\mathrm{TLS}} + U_{\mathrm{torsion}} + U_{\mathrm{bond}} \tag{7.34}$$

where U_{crystal} represents the fact that a crystal itself is generally an anisotropic field that will result in the intensity falling off in an anisotropic manner, U_{TLS} represents a translation/libration/screw (TLS) – that is, the overall motion of molecules or domains, U_{torsion} is the oscillation along torsion angles and U_{bond} is the oscillation along and across bonds. In principle, all of these contributors are highly correlated, and it is difficult to separate them from each other. Never-

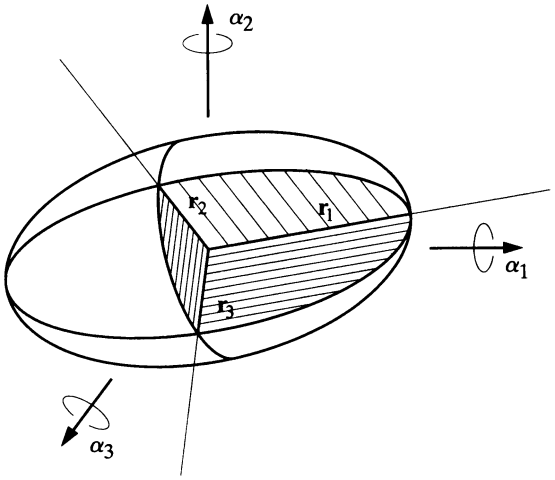


Fig. 7.1 The thermal-ellipsoid model used to represent anisotropic displacement, with major axes indicated. The ellipsoid is drawn with a specific probability of finding an atom inside its contour. Six parameters are necessary to describe the ellipsoid: three represent the dimensions of the major axes and three the orientation of these axes. These six parameters are expressed in terms of a symmetric U tensor whose contribution to atomic scattering is given by Eq. (3.81). (Reproduced with permission from Dauter et al., 2001, International Union of Crystallography.)

theless, the splitting up of U_{atom} into this sum of different contributors makes it possible to apply atomic anisotropic parameters at different resolutions in a distinct way. U_{crystal} and U_{TLS} can be applied at any resolution. U_{crystal} can be regarded as anisotropic overall scale factor and increases the number of parameters by at most a factor of five. U_{TLS} accounts for the anisotropy of the movement of whole molecules or domains and introduces 20 more parameters per molecule or domain. As torsion angles exhibit a strong correlation between each other, they are difficult to model and normally are not separately refined. U_{atom} can only be refined at atomic resolution as defined above, and introduces six parameters per atom. REFMAC (Murshudov et al., 1997), for example, only corrects for U_{crystal} before refining individual atomic anisotropic displacement factors. The derived atomic anisotropy is thus the sum of U_{TLS}, U_{torsion} and U_{atom}.

The full refinement with up to 10 (usually nine) parameters per atom can be made with the classical least-squares residual, as with the programs SHELXL (Sheldrick and Schneider, 1997) and REFMAC (Murshudov et al., 1997) or by maximum-likelihood procedures (Bricogne and Irwin, 1996; Pannu and Read, 1996; Murshudov et al., 1997; REFMAC). The mathematical approaches to minimize the respective residuals have been outlined in the previous sections. The least-squares refinement of large structures may become problematic because the number of terms in the matrix of the normal equations increases with the square of the number of parameters. The full matrix solution becomes unfeasible due to computer time and memory problems. A common approach is then

the block-matrix approximation where, instead of the full matrix, only square blocks along the matrix diagonal are constructed, including groups of parameters that are expected to be correlated. The correlation between parameters belonging to different blocks is therefore neglected completely. In principle, this leads to the same solution, but more slowly and with less precise error estimates. Nevertheless, block-matrix approaches remain essential for tractable matrix inversion for macromolecular structures. Maximum-likelihood methods, when used for full anisotropic refinement, provide results similar to those obtained with least squares, but with improved weights. A remaining limitation is the use of the diagonal approximation, which prevents the computation of standard uncertainties of individual parameters.

It is not very long since the achievement of X-ray diffraction data extending to atomic resolution was a very rare event. However, recent advances in cryogenic techniques, area detectors and the use of synchrotron radiation have enabled macromolecular data to be collected to atomic resolution for an increasing number of proteins (Dauter et al., 1997). Although the importance of a structure determination at atomic resolution is clear, it is extremely significant for the determination of the metal binding site in a metalloprotein. At lower resolution, stereochemical restraints must be applied, and it is very difficult to formulate such restraints for a metal binding site. In the case of atomic resolution, no prior stereochemical knowledge is necessary and metal–ligand bond distances and angles can be achieved directly. This allows, for example, determination of the redox state of the metal because metal–ligand distances differ at various redox states. An example of a complex metal active site determined at atomic resolution (1.1 Å) is shown in Figure 7.2. The figure depicts the bimetallic active site of CO dehydrogenase of the eubacterium *Oligotropha carboxidovans* in its oxidized state (Dobbek et al., 2002). The atoms of the metal ions and some li-

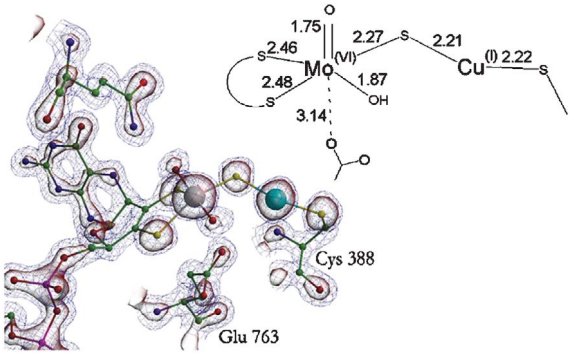


Fig. 7.2 Geometry of the bimetallic active site of CO dehydrogenase of *Oligotropha carboxidovans*. The electron density map has been contoured at 1.0σ. (Reproduced with permission from Dobbek et al., 2002, National Academy of Sciences of USA.)

gands are resolved as spheres in the electron density map. The other atoms show such a resolution that the electron density reduces to more than 50% in the middle between two atoms. A schematic representation of the metal coordination is included as cartoon.

7.3 Verification and Accuracy of Structure Determination

7.3.1 Free *R*-Factor as a Tool for Cross-Validation in Structure Determination

A measure of the quality of a structure determination is the crystallographic *R*-factor given in Eq. (7.22). For a high-resolution structure (e.g., 1.6 Å), the factor should not be much larger than 0.16. As the *R*-factor is an overall number, it does not indicate major local errors; rather, these can be obtained by the evaluation of a real space *R*-factor (Jones et al., 1991) which is calculated on a grid for nonzero elements according to Eq. (7.34):

$$R_{\text{real space}} = \frac{\sum |\rho_{\text{obs}} - \rho_{\text{calc}}|}{\sum |\rho_{\text{obs}} + \rho_{\text{calc}}|} \tag{7.34}$$

where ρ_{obs} is the observed and ρ_{calc} the calculated electron density.

It has been shown that the conventional *R*-factor may reach rather low values in a crystallographic refinement with structural models that were later found to be incorrect. Fortunately, gross errors can be recognized by independent structure solution or by comparison to known structures of homologous macromolecules. However, diffraction data can be misinterpreted in more subtle ways. For example, it is possible to overfit the diffraction data by introducing too many adjustable parameters, or to make the stereochemical restraints too weak. A typical problem arises when too many water molecules are fitted to the diffraction data, thus compensating for errors in the model or the data. A related issue is the overinterpretation of models by placing too much faith in the accuracy of atomic positions at the particular resolution of the diffraction data. The underlying reason for such an overfitting can be found in the relationship between the *R*-factor and the target function for the crystallographic refinement (Eq. 7.19) that is aimed to be minimized. If one assumes that all observations are independent and normally distributed, it can be shown that the target function of Eq. (7.19) is a linear function of the negative logarithm of the likelihood of the atomic model (Press et al., 1986). It transpires that the target function – and thus the *R*-factor – can be made arbitrarily small simply by increasing the number of model parameters used during refinement, regardless of the correctness of the model.

In order to overcome this unsatisfactory situation, Brünger (1992, 1993) proposed the additional calculation of a so-called free *R*-factor, which corresponds to an application of the statistical method of cross-validation (Stone, 1974). Fortu-

nately, single crystal X-ray diffraction data also of biomacromolecules are overdetermined, and in most cases it is possible to set aside a certain fraction of the diffraction data for cross-validation. For this purpose, the reflections are divided into a working set (e.g., 90%) and a test set (e.g., 10%). The reflections in the working set are used in the crystallographic refinement. The free *R*-factor is then calculated with reflections from the test set which were not used for the crystallographic refinement, and is thus unbiased by the refinement process. Free *R*-factors are generally higher than expected (typically 20%, but sometimes above 30%). There exists a high correlation between the free *R*-factor and the accuracy of the atomic model phases. Furthermore, it is empirically related to coordinate error, and thus high free *R*-factors may be caused by a relatively high coordinate error of the model.

7.3.2 Determination of Coordinate Uncertainty

The topic of coordinate uncertainty of structural models of biomacromolecules determined by single crystal X-ray diffraction has been reviewed (Cruickshank, 1999, 2001). Here, we provide some insight into the problem and discuss the magnitudes used to characterize coordinate uncertainty.

In mathematical statistical theory, a distinction is made between the terms accuracy and precision. A single measurement of the magnitude of a quantity differs by error from its unknown true value λ. For a given experimental procedure, the potential results of an experiment define the probability density function $f(x)$ of a random variable. Both the true value λ and the probability density $f(x)$ are unknown. The problem of assessing the *accuracy* of a measurement is thus the double problem of estimating $f(x)$ and of assuming a relationship between $f(x)$ and λ. *Precision* is linked to the function $f(x)$ and its spread.

The problem of which relationship to assume between $f(x)$ and the true value λ is more difficult, including particularly the question of systematic errors. The usual procedure, after correcting for known systematic errors, is to assume that some typical property of $f(x)$, often the mean, is the value of λ. No repetition of the same experiment will ever detect the systematic errors, so statistical estimates of precision consider only random errors. Empirically, systematic errors can be detected only by remeasuring the quantity with another technique. In the case of single crystal X-ray diffraction, this may be the use of synchrotron radiation in place of a conventional X-ray source, a different crystal or crystal form, or data collection at another temperature of the crystal. Nowadays, the term standard uncertainty (s.u.) has replaced the well-established term estimated standard deviation (e.s.d.).

7.3.2.1 Unrestrained Least-Squares Refinement

The normal equations for the model function $M_i(\mathbf{x})$ given by a nonlinear relationship have been quoted in Eq. (7.13). Here, the model function had been developed in a Taylor series. As a model function we use either the structure fac-

tor amplitude F_{obs} or the intensity I_{obs}. It is assumed that we know approximate values $\mathbf{x}_c$ for the parameters $\mathbf{x}$, which deviate from $\mathbf{x}$ by $\mathbf{e}$, and expand $M_i(\mathbf{x})$ as first-order Taylor series. The normal equations for F_{calc} as model function is then given by

$$\sum_i \varepsilon_i \left[\sum_{\mathbf{h}} w(\mathbf{h}) (\partial F_{calc}/\partial x_{c,i})(\partial F_{calc}/\partial x_{c,j}) \right] = \sum_{\mathbf{h}} w(\mathbf{h})(F_{obs} - F_{calc})(\partial F_{calc}/\partial x_{c,j}) \ . \tag{7.35}$$

There are n of these equations for $j = 1, \ldots, n$ to determine the n unknown ε_j.

For $I_{obs} = F_{obs}^2$ and $I_{calc} = F_{calc}^2$ the normal equations are:

$$\sum_i \varepsilon_i \left[\sum_{\mathbf{h}} w(\mathbf{h}) (\partial F_{calc}^2/\partial x_{c,i})(\partial F_{calc}^2/\partial x_{c,j}) \right] = \sum_{\mathbf{h}} w(\mathbf{h})(F_{obs}^2 - F_{calc}^2)(\partial F_{calc}^2/\partial x_{c,j}) \tag{7.36}$$

The index triple $\mathbf{h}$ has been omitted for the structure factor amplitudes in Eqs. (7.35) and (7.36). Both forms of the normal equations may be abbreviated to

$$\sum_i \varepsilon_i a_{ij} = b_j \ . \tag{7.37}$$

Some important points in the derivation of the standard uncertainties of the refined parameters can be most easily understood if we assume that the matrix a_{ij} can be approximated by its diagonal elements. Each parameter is then determined by a single equation of the form:

$$\varepsilon_i \sum_{\mathbf{h}} w(\mathbf{h}) g^2 = \sum_{\mathbf{h}} w(\mathbf{h}) g \Delta \ , \tag{7.38}$$

where $g = \partial F_{calc}/\partial x_{c,i}$ or $\partial F_{calc}^2/\partial x_{c,i}$ and $\Delta = F_{obs} - F_{calc}$ or $F_{obs}^2 - F_{calc}^2$. Hence,

$$\varepsilon_i = \left(\sum_{\mathbf{h}} w(\mathbf{h}) g \Delta \right) \Big/ \sum_{\mathbf{h}} w(\mathbf{h}) g^2 \ . \tag{7.39}$$

At the end of the refinement, when the residual is a minimum the variance (square of s.u.) of the parameter x_i due to the uncertainties is

$$\sigma_i^2 = \left[\sum_{\mathbf{h}} w(\mathbf{h})^2 g^2 \sigma^2(F) \right] \Big/ \left(\sum_{\mathbf{h}} w(\mathbf{h})^2 g^2 \right)^2 \ . \tag{7.40}$$

If the weights have been chosen as $w(\mathbf{h}) = 1/\sigma^2(F_{obs}(\mathbf{h}))$ or $1/\sigma^2(F_{obs}^2(\mathbf{h}))$, this simplifies to

$$\sigma_i^2 = 1 \Big/ \left(\sum_{\mathbf{h}} w(\mathbf{h}) g^2 \right) = 1/a_{ii} \,, \tag{7.41}$$

which is appropriate for absolute weights. Equation (7.41) provides an s.u. for a parameter relative to the s.u.'s $\sigma(F_{\text{obs}}(\mathbf{h}))$ or $\sigma(F_{\text{obs}}^2(\mathbf{h}))$.

In general, with the full matrix a_{ij} in the normal equations,

$$\sigma_i^2 = (a^{-1})_{ii} \,, \tag{7.42}$$

where $(a^{-1})_{ii}$ is an element of the inverse matrix of a_{ij}.

7.3.2.2 Restrained Least-Squares Refinement

The residuals to be minimized in restrained least-squares refinement has been given by Eqs. (7.17) to (7.19) for energy restraints and by Eq. (7.20) for stereochemical restraints. In a high-resolution unrestrained refinement of a small molecule, the s.u. of a bond length A–B is often well approximated by

$$\sigma(l) = (\sigma_A^2 + \sigma_B^2)^{1/2} \,. \tag{7.43}$$

However, in a biological macromolecule determination $\sigma(l)$ is often much smaller than either σ_A or σ_B, due to the excellent information from the stereochemical dictionary, which correlates the positions of A and B.

The determination of the precision of a restrained refinement of a biological macromolecule is, in principle, straightforward. The inverse of the final full matrix delivers estimates of the variances and covariances of all parameters. The dimensions of the matrix are the same for both unrestrained or restrained refinement.

The s.u. for a parameter refined with restraints has been derived for a simple structural model consisting of two bonded atoms to give a flavor of the influence of the restraints on the s.u. The variance of the restrained length is given as

$$1/\sigma_{\text{res}}^2(l) = 1/\sigma_{\text{diff}}^2(l) + 1/\sigma_{\text{geom}}^2(l) \,, \tag{7.44}$$

where $\sigma_{\text{diff}}^2(l)$ is the variance of the diffraction term and $\sigma_{\text{geom}}^2(l)$ that of the geometrical restraint. Thus, the restrained refinement determines a length which is the weighted mean of the diffraction-only length and the geometric dictionary length.

7.3.2.3 Rough Estimation of Coordinate Uncertainties

With the increasing power of computers and more efficient algorithms (e.g., Tronrud, 1999; Murshudov et al., 1999), a final matrix should be computed and inverted more regularly, and not only for high-resolution analyses. However, in

the normal practice refinement programs are used that are unable to deliver estimates of coordinate uncertainties directly. The standard method applied is to obtain an estimation of the mean positional error for the whole structural model. Several approaches can be used, and these will be described briefly in the following sections.

7.3.2.3.1 Luzzati Plot

Luzzati (1952) derived a formula, which determines the *R*-factor as a function of $\langle \Delta r \rangle$, the average radial error of the atomic position. His analysis shows that the *R*-factor is a linear function of $S = 2 \sin \theta / \lambda$ and $\langle \Delta r \rangle$ for a substantial range of $S\langle \Delta r \rangle$, with

$$R(S, \langle \Delta r \rangle) = (2\pi)^2 S \langle \Delta r \rangle \ . \tag{7.45}$$

The theoretical Luzzati plots of the *R*-factor are nearly linear for small-to-medium *S*. If one plots the *R*-factor as a function of *S*, it is possible to determine the actual average radial error $\langle \Delta r \rangle$ by comparing the curve with the theoretical curves given in Eq. (7.45). The determination of $\langle \Delta r \rangle$ by a Luzzati plot has been common in the past. However, as Luzzati plots are very different from other statistical estimates of error, they have been criticized (see Cruickshank, 2001), and the use of other estimates derived by a more suitable statistical analysis has been recommended.

7.3.2.3.2 σ_A-Plot

Read (1986, 1990) proposed another graphical method to obtain the average positional error $\langle \Delta r \rangle$. In this approach, σ_A is plotted as function of *S* in the following form:

$$\ln \sigma_A = \frac{1}{2} \ln \left(\frac{\sum_{i=1}^{P} f_i^2}{\sum_{j=1}^{N} f_j^2} \right) - \frac{\pi^3}{4} \langle \Delta r \rangle^2 S^2 \ , \tag{7.46}$$

where σ_A is defined by Eq. (6.36). The summation goes over all *N* atoms of the whole structure and *P* atoms of the partially known structure. The argument of the logarithm on the right-hand side of Eq. (7.46) should be 1 if the structure refinement has been finished. However, this is never the case because of the disordered structure of the solvent atoms in the crystal. These atoms contribute considerably to low-resolution reflections which should, therefore, be ignored at the final stage of refinement. σ_A plots can, for example, be calculated with program SIGMAA (Read, 1986) or within CNS (Brünger et al., 1998).

7.3.2.3.3 The Diffraction-Component Precision Index

Cruickshank (1999) introduced a quick and rough guide for the diffraction-data-only error component for atoms with isotropic *B*-factor equal to the average *B*-factor, B_{avg}, of the biomacromolecular structure. This is named the diffraction component precision index (DPI), and is given by:

$$\sigma(x, B_{avg}) = 1.0(N_i/p)^{1/2} C^{-1/3} R d_{min} \ , \tag{7.47}$$

where N_i is the number of fully occupied atomic sites, p is the difference between the number of observations n_{obs} and the number of parameters n_{param}, and C is the fractional completeness of the diffraction data to d_{min}, the minimal lattice plane distance.

For low-resolution structures, the number of parameters may exceed the number of diffraction data. In Eq. (7.47), p is then negative, so that $\sigma(x)$ is imaginary, but this problem may be circumvented empirically by replacing p with n_{obs} and R with R_{free}. The counterpart of the DPI (Eq. 7.47) is then

$$\sigma(x, B_{avg}) = 1.0(N_i/n_{obs})^{1/2} C^{-1/3} R_{free} d_{min} \ . \tag{7.48}$$

Here, n_{obs} is the number of the reflections included in the refinement, not the number in the R_{free} set.

Often, an estimate of a position error $\langle \Delta r \rangle$, rather than a coordinate error $\langle \Delta x \rangle$, is required. In the isotropic approximation we obtain

$$\sigma(x, B_{avg}) = 3^{1/2} \sigma(x, B_{avg}) \ . \tag{7.49}$$

Consequently, the DPI formulae for the position errors are

$$\sigma(x, B_{avg}) = 3^{1/2} (N_i/p)^{1/2} C^{-1/3} R d_{min} \tag{7.50}$$

with R and

$$\sigma(x, B_{avg}) = 3^{1/2} (N_i/n_{obs})^{1/2} C^{-1/3} R_{free} d_{min} \tag{7.51}$$

with R_{free}.

7.3.3 Validation of the Geometric and Stereochemical Parameters of the Structural Model

In the previous sections we discussed the attainment of the most reliable structural model of a biomacromolecule and the estimation of its positional uncertainty. Usually, however, the data obtained from the diffraction experiment(s) are not of sufficiently high resolution to define the atomic positions of a macromolecule with adequate precision. In such cases, the crystallographic refinement procedures use additional restraints based on prior knowledge about the chemical structure of the molecule and its conformational properties.

When building the model of a protein structure, we must consider the energy-based rules for the conformation of the polypeptide chain. A polypeptide chain in extended conformation is shown in Figure 7.3. This consists of the single peptide groups which are linked to each other by the peptide bond connecting each main chain C-atom with the adjacent N-atom. For structural properties, it is useful to divide the main chain into repeating units extending from one C_α atom to the next C_α atom. These units are planar and rigid, and are linked into a chain by covalent bonds at the C_α atoms; the only degrees of freedom they have are rotations around these bonds. Each unit can rotate around two of such bonds: the C_α-C bond and C_α-N bond. The angle of rotation around the C_α-N bond is called ϕ, and that around the C_α-N bond is called ψ. Most combinations of ϕ and ψ angles for an amino acid are not allowed because of steric collisions between the side chains and main chain.

The conformation of the main chain folding is verified by a Ramachandran plot (Ramachandran et al., 1963). The dihedral angles ϕ and ψ are plotted against each other for each residue. The data points should lie in the allowed regions of the plot which correspond to energetically favorable secondary structures such as α-helices, β-sheets and defined turn structures. Exceptions are glycine residues, which may occur at any position in the Ramachandran plot. After each round of model building and refinement, a Ramachandran plot should be compiled and, at the final stage, all data points should lie in energetically favored regions of the plot (see Fig. 7.4). Figure 7.4a shows a Ramachandran plot of the initial structure of the small subunit of ribulose-1,5-biphosphate carboxylase/oxygenase (RiBisCO), while Figure 7.4b shows a plot of the refined structure. The initial structure contains several amino acid residues in energetically disallowed regions, which is no longer the case for the refined structure.

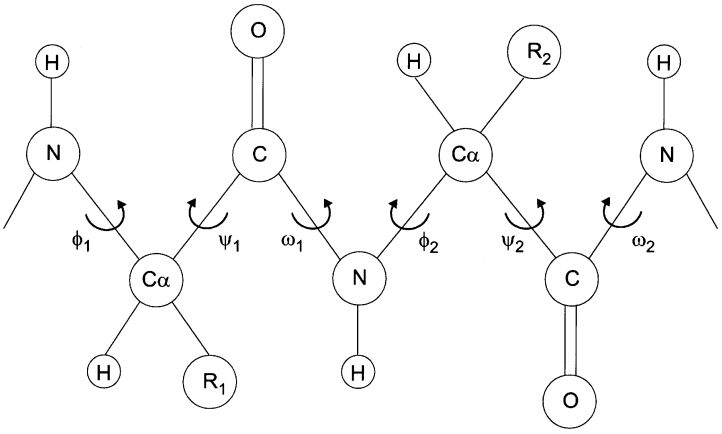


Fig. 7.3 Schematic drawing of a polypeptide chain in fully extended conformation. The meaning of the dihedral angles ϕ and ψ is explained in the text. The peptide planes are usually flat with $\omega = 180^\circ$.

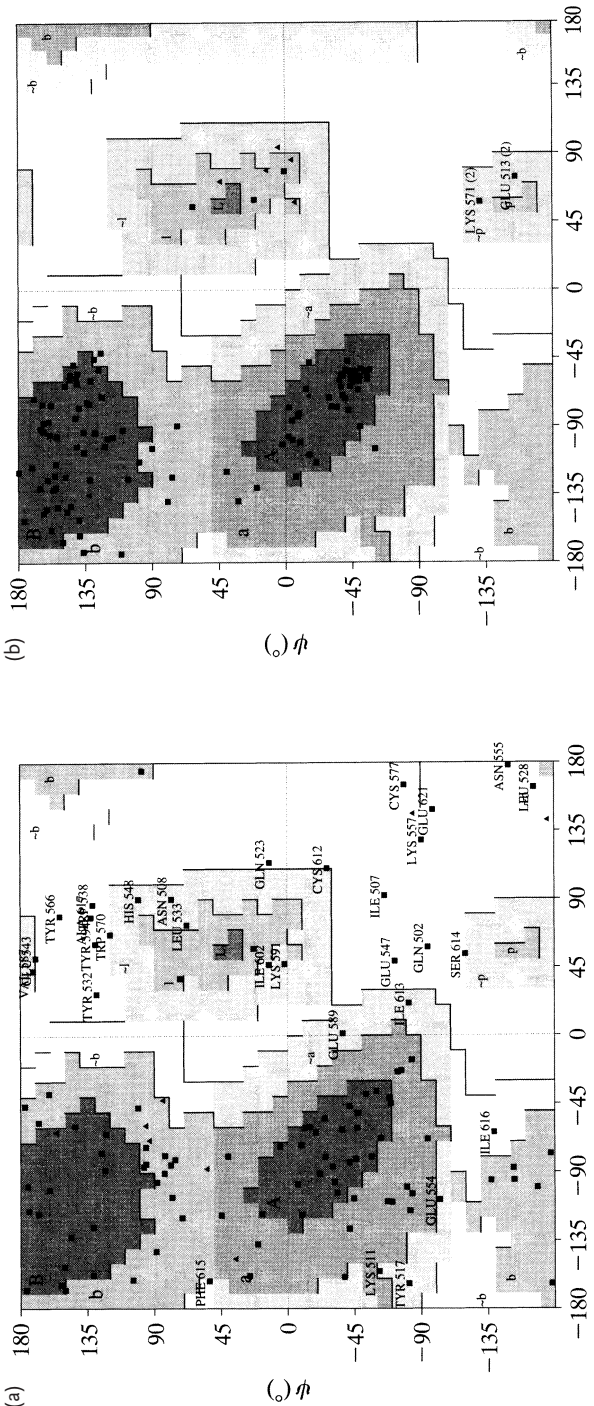


Fig. 7.4 Ramachandran diagrams from PROCHECK (Laskowski et al., 1993) for the small subunit of ribulose-1,5-biphosphate carboxylase/oxygenase (RiBisCO). (a) Initial structure; (b) refined structure. (Reproduced with permission from Dym et al., 2001, International Union of Crystallography.)

The covalent geometry of the atomic structure is checked by comparisons against standard values derived from crystals of small molecules. For proteins, the most commonly used standard values for bond distances and angles are those compiled by Engh and Huber (1991), from molecular fragments in the Cambridge Structural Database, CSD (Allen et al., 1979), that most closely resemble chemical groups in amino acids.

Protein-structure validation packages, such as PROCHECK (Laskowski et al., 1993) or WHAT IF (Hooft et al., 1996) flag all bond distances and angles that deviate significantly from the database-derived reference values. This includes analysis of the deviations from planarity in aromatic rings and planer side-chain groups.

Similar checks are performed for the covalent geometry of the atomic models of RNA or DNA oligo- and polynucleotides. Here, standard ranges for bond distances and angles are derived from crystal structures of nucleic acid bases, mononucleosides and mononucleotides in the CSD (Clowney et al., 1996; Gelbin et al., 1996).

Validation of the covalent geometry of the so-called "hetero groups" (chemically modified monomer groups or small molecules that bind to macromolecules) is much more difficult because reference data bases cannot be created easily due to a strong influence on the conformation of the ligand molecule upon binding to the macromolecule. At present, this is not carried out routinely in the Protein Data Bank (PDB) (Bernstein et al., 1977; Berman et al., 2000) and, as result, the quality of the hetero groups deposited in the PDB varies considerably.

Other stereochemical parameters, such as side-chain torsion angles (χ_1, χ_2, χ_3, etc.), the peptide bond torsion ω, the C_α tetrahedral distortion, disulfide geometry, and non-bonded parameters, such as close van der Waals contacts, geometry of H-bonds and salt bridges and interactions in the solvent structure, must be checked. This can conveniently be done with the program PROCHECK (Laskowski et al., 1993).

7.3.4 Validation of the Structural Model against the Experimental Data

The basic concepts and reliability indices and magnitudes have been discussed and listed in Sections 7.3.1 and 7.3.2. A systematic approach to perform this task at the end of a structure determination has been offered by the program SFCHECK (Vaguine et al., 1999). The program reads in the structure factor data and the PDB-file holding the structural model. It then calculates structure factors from the model and determines the scale factor between them and the observed structure factors. For the global agreement between the model and the experimental data, the R- and R_{free}-factors and the correlation coefficient

$$CC_F = \frac{(F_{obs}F_{calc}) - \langle F_{obs}\rangle\langle F_{calc}\rangle}{\left[\left(\langle F_{obs}^2\rangle - \langle F_{obs}^2\rangle^2\right)\left(\langle F_{calc}^2\rangle - \langle F_{calc}\rangle^2\right)\right]^{1/2}} \quad (7.52)$$

between the calculated and observed structure factor amplitudes are compiled. The estimation of errors in atomic positions is performed by the DPI (Eq. 7.50) and from a Luzzati plot. In addition to the global structure quality measures, SFCHECK also determines the quality of the model in specific regions. Several quality estimators can be calculated for each residue in the macromolecule and, whenever appropriate, for solvent molecules and groups of atoms in ligand molecules. These estimators are the normalized atomic displacement (Shift), the correlation coefficient between calculated and observed electron densities (Density correlation), the local electron-density level (Density index), the average (*B*-factor) and the connectivity index (Connect), which measures the local electron-density level along the molecular backbone. These quantities are computed for individual atoms and averaged over those composing each residue or group of atoms.

7.3.5
Deposition of Structural Data with the Protein Data Bank

Almost all spatial structures of biological macromolecules determined either by X-ray crystallography or nuclear magnetic resonance (NMR) techniques have been – and will be – deposited with the RCSB Protein Data Bank at Rutgers University (Berman et al., 2000). The PDB-file of the final structural model and the structure factor file must be supplied. The data can be submitted online using the ADIT tool available on the RCSB web site (http://www.rcsb.org/). For this, one must supply a row of additional information such as the reference to one or more publications about the structure, biological source, production and crystallization of the macromolecule, and many more details. The information relating to the structural model is in a file that contains, for each individual atom of the model, a record with atom number, atom name, residue type, residue name, coordinates *x*, *y*, *z*, *B*-value(s), and occupancy. The header records hold useful information such as crystal parameters, amino acid sequence, secondary structure assignments, and references.

References

Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R., Watson, D. G., *Acta Crystallogr.* **1979**, *B35*, 2331–2339.

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., Haak, J. R., *J. Chem. Phys.* **1984**, *81*, 3684–3690.

Bernstein, F. C., Keitzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T.,

Tasumi, M., *J. Mol. Biol.* **1977**, *112*, 535–542.
Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindlyalov, I.N., Bourne, P.E., *Nucleic Acids Res.* **2000**, *28*, 235–242.
Bricogne, G., Irwin, J., Maximum-likelihood Structure Refinement: Theory and Implementation within BUSTER+TNT. In: Dodson, E., Moore, M., Ralph, A., Bailey, S. (Eds.), *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, pp. 85–92. Daresbury Laboratory, Warrington, **1996**.
Brünger, A.T., *Nature* **1992**, *355*, 472–475.
Brünger, A.T., *Acta Crystallogr.* **1993**, *D49*, 24–36.
Brünger, A.T., Kuriyan, J., Karplus, M., *Science* **1987**, *235*, 458–460.
Brünger, A.T., Nilges, M., *Q. Rev. Biophys.* **1993**, *26*, 49–125.
Brünger, A.T., Adams, P.D., Clore, G.M., Delano, W.L., Gros, P., Grossekunstleve, R.W., et al., *Acta Crystallogr.* **1998**, *D54*, 905–921.
Clowney, L., Jain, S.C., Srinivasin, A.R., Westbrook, J., Olson, W.K., Berman, H.M., *J. Am. Chem. Soc.* **1996**, *118*, 509–518.
Cruickshank, D.W.J., *Acta Crystallogr.* **1999**, *D55*, 583–601.
Cruickshank, D.W.J., Coordinate uncertainty. In: Rossmann, M.G., Arnold, E. (Eds.), *International Tables for Crystallography*, Vol F, pp. 403–414. Kluwer Academic Publishers, Dordrecht, **2001**.
Dauter, Z., Lamzin, V.S., Wilson, K.S., *Curr. Opin. Struct. Biol.* **1997**, *7*, 681–688.
Dauter, Z., Murshudov, G.N., Wilson, K.S., Refinement at Atomic Resolution. In: Rossmann, M.G., Arnold, E. (Eds.), *International Tables for Crystallography*, Vol. F, pp. 393–402. Kluwer Academic Publishers, Dordrecht, **2001**.
Dennis, J.E., Schnabel, R.E., *Numerical Methods for Unconstrained Optimization and nonlinear Equations*, Prentice-Hall, Englewood Cliffs, **1983**.
Dobbek, H., Gremer, L., Kiefersauer, R., Huber, R., Meyer, O., *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 15971–15976.
Dym, O., Eisenberg, D., Yeates, T.O., Detection of errors in protein models. In: Rossmann, M.G., Arnold, E. (Eds.), *International Tables for Crystallography*, Vol. F, pp. 520–525. Kluwer Academic Publishers, Dordrecht, **2001**.
Emsley, P., Cowtan, K., *Acta Crystallogr.* **2004**, *D60*, 2126–2132.
Engh, R.A., Huber, R., *Acta Crystallogr.* **1991**, *A47*, 392–400.
Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-H., Olsen, W.K., Berman, H.M., *J. Am. Chem. Soc.* **1996**, *118*, 519–529.
Greer, J., *J. Mol. Biol.* **1974**, *82*, 279–301.
Goldstein, H., *Classical Mechanics*. Addison-Wesley, Reading, MA, **1980**.
Hendrickson, W.A., *Methods Enzymol.* **1985**, *115*, 252–270.
Hooft, R.W.W., Vriend, G., Sander, C., Abola, E.E., *Nature* **1996**, *381*, 272.
Jack, A., Levitt, M., *Acta Crystallogr.* **1978**, *A34*, 931–935.
Jones, T.A., *J. Appl. Crystallogr.* **1978**, *15*, 24–31.
Jones, T.A., Thirup, S., *EMBO J.* **1986**, *5*, 819–822.
Jones, T.A., Zou, J.Y., Cowan, S.W., Kjeldgaard, M., *Acta Crystallogr.* **1991**, *A47*, 110–119.
Kleywegt, G.J., *Acta Crystallogr.* **1996**, *D52*, 842–857.
Kleywegt, G.J., Read, R.J., *Structure* **1997**, *5*, 1557–1569.
Kleywegt, G.J., Zou, J.-Y., Kjeldgaard, M., Jones, T.A., Around, O. In: Rossmann, M.G., Arnold, E. (Eds.), *International Tables for Crystallography*, Vol. F, pp. 353–356. Kluwer Academic Publishers, Dordrecht, **2001**.
Konnert, J.H., Hendrickson, W.A., *Acta Crystallogr.* **1980**, *A36*, 344–350.
Laskowski, R.A., MacArthur, M.W., Moss, D.S., Thornton, J.M., *J. Appl. Crystallogr.* **1993**, *26*, 283–291.
Lamzin, V.S., Perrakis, A., Wilson, K.S., The ARP/wARP suite for automated construction and refinement of protein models. In: Rossmann, M.G., Arnold, E. (Eds.), *International Tables for Crystallography*, Vol. F, pp. 720–722. Kluwer Academic Publishers, Dordrecht, **2001**.
Luzzati, V., *Acta Crystallogr.* **1952**, *5*, 802–810.
Murshudov, G.N., Vagin, A.A., Dodson, E.J., *Acta Crystallogr.* **1997**, *D53*, 240–255.

Murshudov, G. N., Vagin, A. A., Lebedev, A., Wilson, K. S., Dodson, E. J., *Acta Crystallogr.* **1999**, *D55*, 247–255.
Pannu, N. S., Read, R. J., *Acta Crystallogr.* **1996**, *A52*, 659–668.
Pannu, N. S., Murshudov, G. N., Dodson, E. J., Read, R. J., *Acta Crystallogr.* **1998**, *D54*, 1285–1294.
Press, W. H., Flannery, B. P., Teukolosky, S. A., Vetterling, W. T., *Numerical Recipes*, Cambridge University Press, Cambridge, **1986**.
Prince, E., Boggs, P. T., Least squares. In: Wilson, A. J. C., Prince, E. (Eds.), *International Tables for Crystallography*, Vol. C, pp. 594–604. Kluwer Academic Publishers, Dordrecht, **1999**.
Prince, E., Finger, L. W., Konnert, J. H., Constraints and restraints in refinement. In: Wilson, A. J. C., Prince, E. (Eds.), *International Tables for Crystallography*, Vol. C, pp. 609–617. Kluwer Academic Publishers, Dordrecht, **1999**.
Ramachandran, G. N., Ramakrishnan, C., Sasisekharan, V. J., *J. Mol. Biol.* **1963**, *7*, 95–99.
Read, R. J., *Acta Crystallogr.* **1986**, *A42*, 140–149.
Read, R. J., *Acta Crystallogr.* **1990**, *A46*, 900–912.
Rice, L. M., Brünger, A. T., *Proteins Struct. Funct. Genet.* **1994**, *19*, 277–290.
Rice, L. M., Shamoo, Y., Brünger, A. T., *J. Appl. Crystallogr.* **1998**, *31*, 798–805.
Roversi, P., Blanc, E., Vonrhein, C., Evans, G., Bricogne, G., *Acta Crystallogr.* **2000**, *D56*, 1313–1323.
Roussel, A., Cambillau, C., Turbo-Frodo in Silicon Graphics Geometry, *Partners Directory*. Silicon Graphics, Mountain View, **1989**.
Schomaker, V., Waser, J., Marsh, R. E., Bergmann, G., *Acta Crystallogr.* **1959**, *12*, 600–604.
Sheldrick, G. M., *Acta Crystallogr.* **1990**, *A46*, 467–473.
Sheldrick, G. M., Schneider, T. R., *Methods Enzymol.* **1997**, 319–343.
Stone, M., *J. R. Stat. Soc. Ser. B* **1974**, *36*, 111–147.
Tronrud, D. E., Ten Eyck, L. F., Matthews, B. W., *Acta Crystallogr.* **1987**, *A43*, 489–501.
Tronrud, D. E., *Acta Crystallogr.* **1999**, *A55*, 700–703.
Ten Eyck, L. F., Weaver, L. H., Matthews, B. W., *Acta Crystallogr.* **1976**, *A32*, 349–350.
Terwilliger, T. C., *Acta Crystallogr.* **2002**, *D59*, 34–44.
Turk, D., MAIN: A Computer Program for Macromolecular Crystallographers, Now with Utilities You Always Wanted. In: *American Crystallography Association Annual Meeting*, Vol. 27 (ISSN 0596-4221), p. 54, Abstract 2m.6.B, **1995**.
Vaguine, A. A., Richelle, J., Wodak, S. J., *Acta Crystallogr.* **1999**, *D55*, 191–205.
Vellieux, F. M. D. A. P., Hunt, J. F., Roy, S., Read, R. J., *J. Appl. Crystallogr.* **1995**, *28*, 347–351.

8
Crystal Structure Determination of the Time-Course of Reactions and of Unstable Species

8.1
Introduction

So far, we have discussed the conventional X-ray experiment, in which we obtain a time-average of the biomacromolecular crystal structure. At first view it appears that, by using a crystal, the structural characterization of the time-course or of unstable species of a chemical or photo-chemical reaction is impossible. This may be true for crystals of small molecules in which the crystal packing does not allow transport of the reactant(s) to the molecules. However, photo-chemical reactions may be an exception. The situation is different for proteins, which constitute the major class of biomacromolecules. Protein crystals may be regarded as concentrated solutions (typically 30–50 mM protein) because of their large solvent content of, usually, 30–80%. The molecules are held in the crystal lattice by relatively few weak interactions, allowing for some flexibility and motion. In many cases, a protein crystal contains solvent channels that allow easy access to the protein by the reactant(s), and subsequent release of the reaction product(s). Crystalline proteins or enzymes are, therefore, often biochemically active (Rossi, 1992). One notable difference from working with concentrated solutions, however, stems from the arrangement of the solvent in ordered arrays of solvent channels. Thus, convection is impossible, and diffusion is restricted in space. Reaction initiation by simple mixing of reactants, as occurs in the stopped- or quenched-flow methods for rapid kinetic experiments in solution, will in general take too long. Also, enzymatic inactivity or reduced activity in the crystalline state may be due to steric restrictions, such as inaccessibility of the active sites or to an inhibition of substrate binding or catalysis caused by the crystallization conditions, such as high salt concentration or unfavorable pH in the crystal. Therefore, the kinetics of the process under investigation have also to be measured in the crystal. This information is also essential for knowing when to collect the diffraction data of an intermediate when the reaction has been initiated in the crystal. In the ideal case, the kinetics in the crystal should be analyzed spectroscopically and noninvasively *in situ*, thereby adding the advantage that this analysis may also be carried out simultaneously with the collection of the X-ray diffraction data.

The prerequisites for time-resolved studies are that the reaction to be studied can be initiated uniformly in space and time throughout the crystal, and that

the time necessary for triggering the reaction and collecting the data is much shorter than the typical lifetimes of the reaction intermediates to be characterized. Intermediate states can be monitored only if they are sufficiently occupied and are stable for the data collection time.

The crystal structures of reactive unstable species, as occurring during chemical reactions, can be determined by time-resolved crystallography or trapping approaches. As discussed earlier, the reaction must be initiated, and for this purpose several triggering methods are available (Schlichting and Goody, 1997; Schlichting, 2000). Both, these methods and the trapping approaches will be briefly outlined in the following sections.

Time-resolved X-ray crystallography is carried out via Laue diffraction experiments. Although Laue diffraction, in which a stationary crystal is illuminated by a polychromatic beam of X-rays, was the original crystallographic technique, it was largely replaced during the 1930s by rotating crystal, monochromatic techniques. With the advent of naturally polychromatic synchrotron X-ray sources and possible short exposure times of $\sim$100 ps (the duration of a single X-ray pulse at a third-generation synchrotron source; Bourgeois et al., 1996), the Laue technique could be used for time-resolved X-ray crystal structure determinations. The Laue technique and its applications have been the subject of several reviews (Clifton et al., 1997; Moffat, 1997, 2001; Schlichting and Goody, 1997; Stoddard, 1998; Ren et al., 1999; Schlichting, 2000; Schmidt et al., 2005). A brief outline of the Laue technique is also presented in the following sections.

8.2 Triggering Methods

Reactions can be initiated by changing thermodynamic parameters such as temperature or pressure, by irradiating with light or other radiation, or by changing the concentration of substrates, cofactors, protons, and electrons. Concentration jumps can be generated most easily by diffusion, but rapid reactions require other approaches. The choice of the trigger depends largely on the physico-chemical properties of the system and the reaction studied. An ideal starting point for time-resolved protein crystallography is provided by the presence of a "built-in" trigger. This may be the case for proteins involved in the conversion of light into other energy forms (e.g., photosynthetic reaction centers), in the transduction of light signals (e.g., bacteriorhodopsin), or that have light-sensitive bonds (e.g., carbon monoxide complexes of heme proteins). The influence of the rapidity of the trigger in relation to the time constants of the system on the detection of intermediates has been demonstrated exemplarily for the photoreactive yellow protein (Ren et al., 1999), which belongs to the systems with built-in triggers, as do light-sensitive carbon monoxide complexes of heme proteins (Schlichting et al., 1994; Šrajer et al., 1996), photosynthetic reaction centers (Stowell et al., 1997), and bacteriorhodopsin (Edman et al., 1999). Triggering enzymatic reactions is usually less straightforward and may require several strategies for capturing different steps.

There are three processes, which can be used to initiate an enzymatic reaction: (i) photolysis; (ii) diffusion; and (iii) radiolysis.

8.2.1 Photolysis

Photolysis is the cleavage of a molecule by irradiation with light, and is ideal for the initiation of a reaction because it may be accomplished very rapidly, depending on the respective (photo)chemistry. Furthermore, photolysis is broadly applicable experimentally. It can be used for crystals mounted in a capillary, in a loop of a humidity control device at ambient temperatures, or in a loop at cryogenic temperatures. Systems that are not inherently light-sensitive can be rendered so by chemically attaching photosensitive, biochemically inactivating groups to, for example, substrates, cofactors or catalytically important residues of the protein (Schlichting, 2000 and references therein). Such "caged compounds" render the system light-sensitive and biologically inert. Commonly used "cage" groups are substituted-2-nitrobenzyls such as 2-nitrophenylethyl (2NPE) that can be cleaved with light of ca. 350 nm wavelength with concomitant production of a nitroso-ketone. An important application of 2NPE groups is in the caging of nucleotides such as ATP and GTP, and it has been used in relation with GTP in a time-resolved Laue study on the Ras protein (Schlichting et al., 1990). Other cage groups successfully used in crystallographic studies are 3,4-dinitrophenyl (attached to phosphate in glycogen phosphorylase b; Duke et al., 1994) and (4,5-dimethoxy-2-nitrophenyl)ethyl or *α*-carboxy-2-nitrobenzyl attached to NADP in a study of isocitrate dehydrogenase (Cohen et al., 1997).

8.2.2 Diffusion

Diffusion is an experimentally straightforward approach to generate concentration jumps of substrates, cofactors, protons, etc. Because of the intrinsic creation of gradients and the competing effects of diffusion and catalysis, reaction initiation by diffusion is suitable only for very slow processes (half-lives of minutes for the rate-limiting species). Typical diffusion times across 200 μm-thick crystals are seconds to minutes, depending on the size of the compound and the solvent channels and the viscosity of the mother liquor. pH changes, if tolerated by the crystal lattice, can be used to trap intermediates (Verschueren et al., 1993) or to initiate single-turnover reactions to be followed by time-resolved crystallography (Singer et al., 1993) using a flow cell (Petsko, 1985). This set-up can also be used for the structure determination of intermediates accumulating under steady-state conditions of the reaction (turnover rates of up to 0.1 s^{-1}). Depending on the solvent (water versus, e.g., 70% methanol), flow cells can be used at ambient and cryogenic temperatures. The study of Fülop et al. (1994), on cytochrome *c* peroxidase, is an excellent example of reaction initiation by diffusion of substrate used in a time-resolved study.

8.2.3
Radiolysis

Radiolysis caused by the interaction of X-rays with matter may occur in synchrotron radiation experiments. Its effect depends on the energy, and thus on the wavelength and intensity, of the X-rays used. Due to the radiation's high intensity, there will be absorption of the radiation, and this may lead to heating and radiation damage of the sample. The latter involves the generation of photoelectrons that recombine with water to form hydrated electrons, leading to a range of subsequent radical reactions. The absorption will be increased if the protein contains metals and the wavelength of the used X-rays is close to an absorption edge of the metal. X-ray-induced reduction has been observed in many metal-containing systems, and has been used deliberately in experiments on cytochrome P450 (Schlichting et al., 2000). As X-ray absorption is strongly wavelength-dependent, reduction can be minimized by using very short X-ray wavelengths. Increasing the wavelength can be used to generate photoelectrons that may reduce the system under investigation, thereby initiating a reaction (Schlichting et al., 2000).

8.3
Trapping Methods

In the previous section we have described the methods to initiate the reaction. There are now two possible ways to perform the respective X-ray experiment: (i) by trapping the unstable intermediate; or (ii) by accomplishing a time-resolved Laue diffraction experiment. Here, we will briefly explain existing trapping methods. The techniques used to extend the lifetime of an intermediate fall into two broad categories, namely physical and chemical trapping.

8.3.1
Physical Trapping

The rate constant of a single step in a reaction decreases at lower temperatures, with the magnitude of decrease being dependent on the absolute activation energy. For a step with an energy barrier of 4 k_{cal} mol^{-1}, a rate constant measured at 20 °C can be up to 40-fold lower at –80 °C; for an energy barrier of 10 k_{cal} mol^{-1}, the same rate constant can be diminished by over 10^4. By rational lowering of the temperature of the crystal during turnover, it may be possible to trap and observe a rate-limiting species ("freeze trap"). Using this approach, the total energy available to the system is decreased substantially. The clear advantage is an increase in the lifetime of the intermediate of interest which may, under appropriate conditions, become virtually infinite, thereby allowing X-ray data collection with sufficiently long exposure times.

Flash-cooling (as described in Section 2.3.2) can be used for intermediate trapping. In such experiments, the intermediate species accumulates at a phys-

iologically relevant temperature in response to natural rate barriers. The reaction or turnover event is then rapidly quenched to cryo-temperatures for data collection ("trap freeze"). Under these conditions, protein structures experience a lowering of mobility and flexibility that is similar to a phase transition. This effect hinders both the dynamic freedom of the protein and the free exchange of solvent necessary for reactivity.

8.3.2 Chemical Trapping

The population and relative occupancy of a specific catalytic intermediate may be elevated and its structure determined by adjusting the reaction conditions so that the respective intermediate has a lower free-energy than any other state. In essence, the free-energy profile of the catalytic reaction is changed and exploited in order to impose a novel kinetic rate limit or a thermodynamic dead end. Such techniques can be used to isolate either an intermediate within the context of a single turnover experiment or a high-occupancy, steady-state complex during the multi-turnover protocol. Such experiments may incorporate either a significant change to the pH of the reaction, or a perturbed or even nonaqueous mother liquor (Yennewar et al., 1994). Alternatively, enzymes that catalyze single-substrate/single-product reactions (or that proceed through separable half-reactions) may be studied under conditions of thermodynamic equilibrium that favor a single predominant species. Yet another method of chemical trapping is the use of site-directed mutagenesis to create a system for a specific catalytic intermediate (Bolduc et al., 1995).

8.4 Laue Diffraction

8.4.1 Principles of the Laue Technique

A Laue diffraction pattern is obtained when a stationary crystal is illuminated by a polychromatic X-ray beam spanning the wavelength range from λ_{min} to λ_{max}, the so-called band pass. Figure 8.1 shows the Ewald construction for the Laue technique. A reciprocal lattice point that lies in the hatched area (the region between the limiting Ewald spheres of radii $1/\lambda_{min}$ and $1/\lambda_{max}$ and within a sphere of radius $1/d_{min}$, the limiting resolution of the crystal) is in diffraction position for the wavelength λ and will contribute to a spot on the Laue diffraction pattern. A Laue pattern may be thought of as the superposition of a series of monochromatic still diffraction patterns, each taken at a different wavelength of X-rays – that is, a contraction of the Ewald spheres with radii from $1/\lambda_{min}$ to $1/\lambda_{max}$. However, it is more advisable to stretch the reciprocal lattice because the mutual position of the sphere and the lattice is essential to determine the direction of the reflected beam. This

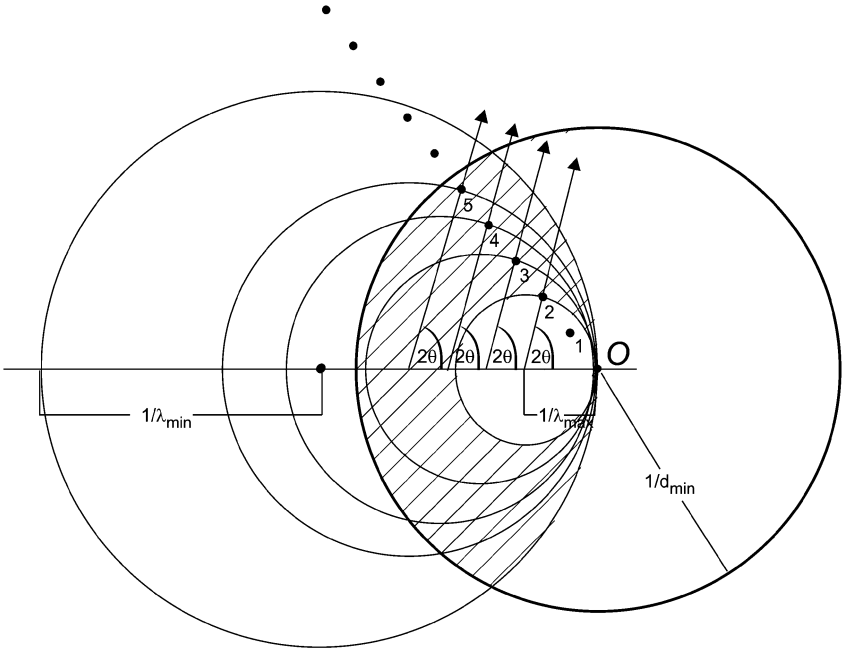


Fig. 8.1 Ewald construction for the Laue diffraction technique. The largest sphere has a radius of $1/\lambda_{min}$ and the smallest one a radius of $1/\lambda_{max}$. The resolution sphere exhibits a radius of $1/d_{min}$. Reciprocal lattice points lying in the hatched area give rise to reflections. The generation of multiple spots is illustrated.

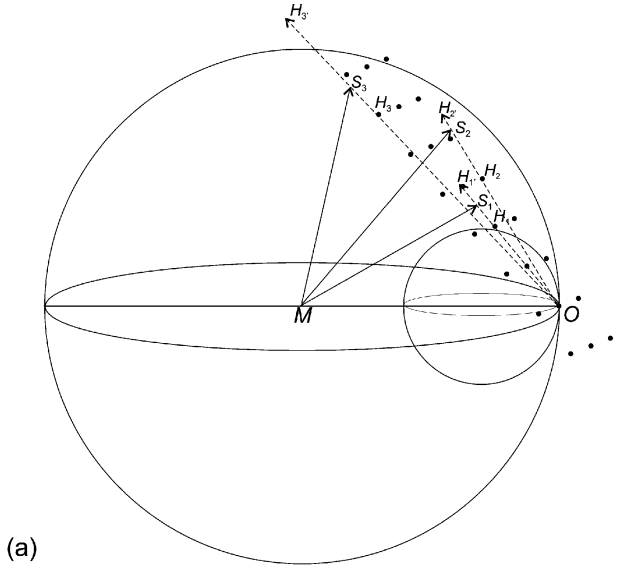
approach is illustrated in Figure 8.2. Each lattice point H moves on the line OH to the point H' with a distance from O of (Eq. 8.1)

$$OH' = OH\frac{\lambda_{max}}{\lambda_{min}} \ . \tag{8.1}$$

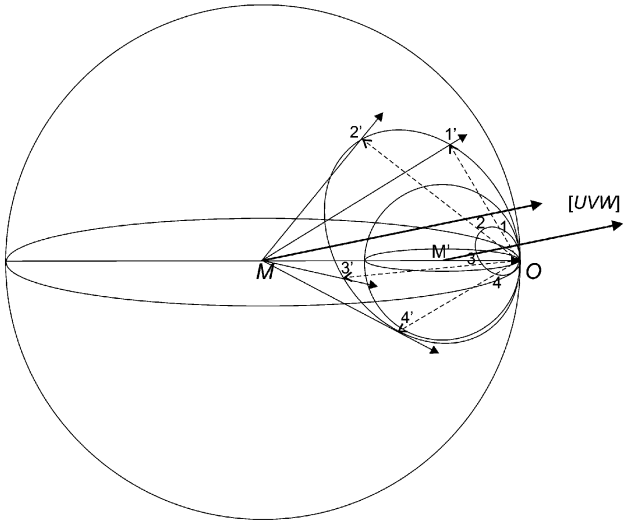
This is carried out in Figure 8.2a for the points H_1 to H_3. Each point H between the spheres for λ_{min} and λ_{max} and no other point intersects the largest sphere in a point S, and MS indicates the direction of the reflected beam whose indices coincide with those of H; for its wavelength, the following relationship holds,

$$\frac{OS}{OH} = \frac{\lambda}{\lambda_{min}} \ . \tag{8.2}$$

This means that Laue spots arise from mapping of rays (lines emanating from the origin in reciprocal space) onto the detector, in contrast to spots in a monochromatic pattern that arise from mapping of individual reciprocal lattice points. Because a ray may contain only one reciprocal lattice point H with indices ($h\ \ k\ \ l$) or several ($h\,k\,l, 2h\,2k\,2l, \ldots, nh\,nk\,nl$) between the origin and



(a)



(b)

Fig. 8.2 Explanation of the geometry of Laue diffraction patterns. (a) Expansion of reciprocal lattice by rays along the lines between the origin of the reciprocal lattice *O* and the respective reciprocal lattice points. (b) Expansion of a circle resulting from the intersection of a reciprocal lattice plane passing through origin *O* with the smallest Ewald sphere. The expansion is illustrated by four rays and the expanded circle, which cuts the largest sphere, as shown. The projection of this circle from the center *M* of the largest Ewald sphere reveals a cone with cone axis [*UVW*]. The directions of reflected beams lie on the surface of the cone.

$1/d_{min}$, a Laue spot may be single, arising from only one reciprocal lattice point, wavelength, and structure factor, or multiple, arising from several points. If a spot is multiple, its integrated intensity is the sum of the integrated intensities of each component. The generation of multiple spots is also illustrated in Figure 8.1, where the reciprocal lattice points 2 to 5 from the indicated reciprocal lattice line produce a multiple Laue spot on the detector.

We consider planes of the reciprocal lattice passing the origin of this lattice, which do not intersect the largest reflection sphere. As an example, we consider a plane that intersects the smallest sphere, as depicted in Figure 8.2 b, and cuts this sphere in a circle. Reciprocal lattice points lying on this circle would give rise to reflections in the directions of a cone with M' as apex, including the direction $M'O$. Its axis $[UVW]$ is normal to the reciprocal lattice plane. If we expand this circle, which is part of the respective reciprocal lattice plane according to Eq. (8.1), the circle expands along the rays until it cuts the largest reflection sphere, which when projected from M yields a cone containing the direction MO with the same zone axis $[UVW]$ (Fig. 8.2 b). This behavior is valid for all reciprocal lattice planes passing the origin of the lattice. Therefore, the spots lie on conic sections, each corresponding to a zone $[UVW]$. These cones intersect a detector normal to the incident beam in ellipses or elliptic arcs designated as "lunes". This means that all reciprocal lattice points of a plane passing through the origin of the reciprocal lattice and lying between the largest and smallest Ewald sphere and the resolution sphere map into the corresponding cone on the largest Ewald sphere. A typical Laue diffraction image plus its predicted diagram is shown in Figure 8.3. Prominent nodal spots, surrounded by clear areas devoid of spots, lie at the intersection of well-populated zones and correspond to rays whose inner point of low, coprime, indices $(h\ k\ l)$. The nodal spots always are multiple. Each single spot is characterized by a unique wavelength λ, associated with the Ewald sphere on which the reflection lies. Extraction of structure amplitudes from single Laue spots requires the derivation and application of a wavelength-dependent correction factor known as the wavelength-normalization curve or λ curve, the value of which varies from reflection to reflection.

An important question here is what fraction of the reciprocal lattice can be registered in a given single Laue experiment. The number of possible observed reflections N_{Laue} will be the quotient of the volume between the Ewald spheres for λ_{min} and λ_{max} and the resolution sphere and the volume of the unit cell of the reciprocal lattice. We quote the result here as

$$N_{Laue} = \frac{1}{4d_{min}^4 V^*}(\lambda_{max} - \lambda_{min}) \ . \tag{8.3}$$

The complete volume stimulated in a single Laue exposure usually is much larger than in a typical monochromatic rotation exposure, and N_{Laue} can be large, particularly for crystals that diffract to high resolution. Hence, Laue diffraction patterns often are crowded, and spatial overlaps between adjacent spots

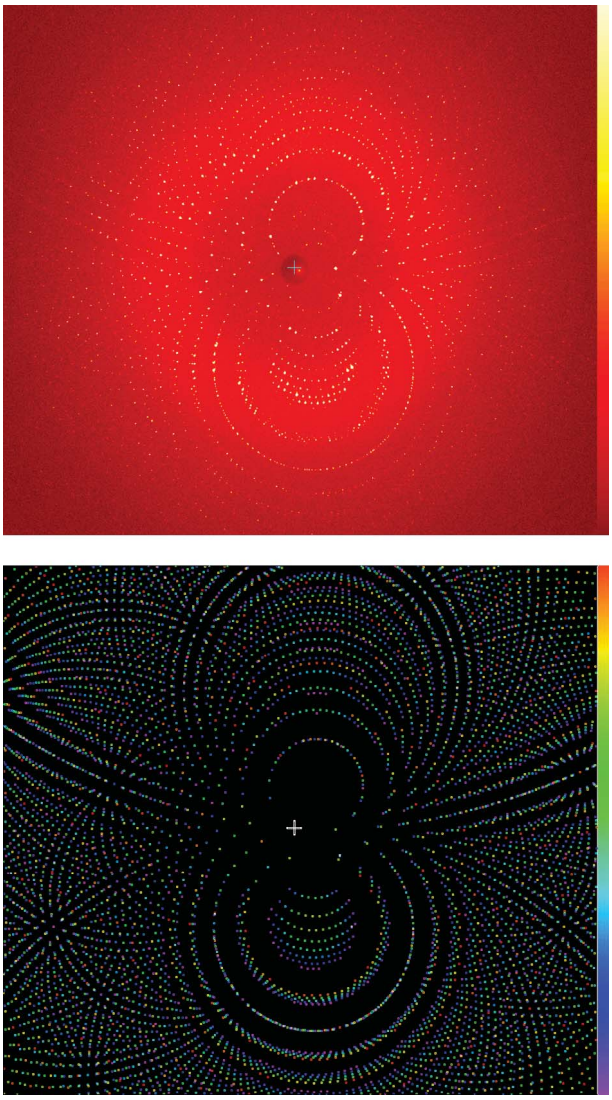


Fig. 8.3 (a) Laue diffraction image of a crystal of the L29W mutant of sperm whale myoglobin. Color code: increasing intensity from red to yellow; data collected at Advanced Photon Source (APS), Argonne National Laboratory, Argonne/USA, beamline 14ID-B, exposure time 40×100 ps after initiation of reaction, bandpass, 1.04 to 1.4 Å (maximum at 1.1 Å). (b) Calculated diffraction pattern at the same crystal orientation. Color code: from purple, higher energy of X-rays or smaller wavelength to red, lower energy of X-rays or larger wavelength. (Source: Dr. Marius Schmidt, Technische Universität München, Physik Department, München, Germany.)

may be common. Nevertheless, a single Laue exposure yields a completeness of diffraction data that is far from being sufficient. This is due to the resolution hole at low resolutions, because all reciprocal lattice points within the smallest Ewald sphere (Fig. 8.1) give no rise for reflections, and low-resolution reflections can be obtained only from a very narrow region. Furthermore, for a typical Laue experiment with $(\lambda_{max} - \lambda_{min}) = 1$ Å and $d_{min} = 2$ Å, the corresponding rotation angle in the monochromatic rotation technique would be about 17°, which relates to a low data completeness, even in crystals of higher symmetry. This means that Laue exposures from several crystal orientations (usually 3 to 10) must be taken to collect a complete and highly redundant data set. A significant redundancy is also necessary for the derivation of an accurate wavelength-normalization curve from the Laue intensities (Helliwell et al., 1989).

The Laue experiment affords a significantly lower exposure time than the monochromatic rotation experiment. The ratio for the Laue exposure time Δt_L and the monochromatic rotation exposure time Δt_M (for a derivation, see Moffat, 1997) is given by

$$\Delta t_L / \Delta t_M = \frac{\tan\theta}{a} \cdot \frac{\Delta k}{k} \ , \tag{8.4}$$

where k is the wave vector. The bandpass $\Delta k/k$ typically has a value $\leq 2 \times 10^{-4}$. If we assume realistic values of $\tan\theta = 0.1$ and $a = 2^\circ = 0.035$ rad, then $\Delta t_L / \Delta t_M \approx 6 \times 10^{-4}$. For this typical case, the Laue exposure time is thus between three and four orders of magnitude less than the corresponding monochromatic exposure time. This makes it the method of choice for time-resolved crystallographic studies.

Crucial to quantitative application of the Laue method is the fact that each reflection is measured at a different wavelength. The square of the structure factor amplitude is related to the measured intensity I of the Laue reflection by

$$|F|^2 = I[Kg(\theta)j(x)f(\lambda)]^{-1} \ , \tag{8.5}$$

where values of K and the functions $g(\theta)$ and $j(x)$ are generally known. Hence, the extraction of the square of the structure factor amplitudes from measured intensities depends on the knowledge of the function $f(\lambda)$, the wavelength-normalization curve. This curve may be derived by one of several methods. At present, the usual method is by examination of redundant measurements of the same reflection, or its symmetry mates, that are stimulated by different wavelengths at different crystal orientations (Campbell et al., 1986; Helliwell et al., 1989; Ren and Moffat, 1995 a).

8.4.2
Advantages and Disadvantages

The Laue technique offers the following advantages:

1. The shortest possible exposure time, making it the method of choice for time-resolved X-ray structure determinations that demand high time resolution.
2. No partially recorded reflections; all spots in a local region of detector space have an identical profile.
3. A large volume of reciprocal space is stimulated in a single exposure; hence only a few exposures at different crystal settings may yield a data set of the necessary completeness.
4. A greater coverage of reciprocal space, and hence greater redundancy and completeness, is achieved automatically at higher resolution where the intensities are naturally weaker.

On the other hand, some disadvantages must also be noted:

1. Multiple Laue spots, activated at several wavelengths (energy overlaps) must be resolved into their component single spots to obtain Laue data sets with sufficiently high completeness.
2. The already mentioned "low-resolution hole" causes a low completeness in the low-resolution region. Very often, there are strong reflections at low resolution. If these are missing, Laue-derived electron density maps will be disturbed due to series termination errors.
3. The exposure of a crystal to an intense polychromatic Laue beam may cause severe radiation damage, with resulting disorder of the crystal.
4. The wide wavelength bandpass activates more reflections, but at the expense of an increased background underlying these spots.
5. The Laue technique is inherently sensitive to crystal disorder, which increases the mosaicity and thus the volume of a reciprocal lattice point. Whatever its origin, it increases spatial overlaps and reduces the peak intensity value of the spot, thus lowering the accuracy of the intensity data set.

In summary, the Laue technique is the method of choice for rapid time-resolved X-ray crystal structure determinations. Data sets from several crystal orientations must be collected to receive a data set with the required completeness and quality. Its application will be limited to crystals of low mosaicity and smaller unit cells in order to avoid the negative effect of overlapping reflections.

8.4.3
Practical Aspects

Rapid time-resolved Laue diffraction requires some special design of the synchrotron radiation beamline used. The bandpass $(\lambda_{max} - \lambda_{min})$ is generated by a bending magnet or wiggler in almost all cases, which provides a broad, smooth

spectrum extending from the beryllium window cut-off 6 keV (2 Å) up to roughly three times the critical energy of the source. The following adjustments are recommended (Moffat, 1997) to collect Laue diffraction data sets of high quality:

- adjust λ_{min} to 0.5 Å or to the wavelength corresponding to three times the critical energy of the source;
- adjust λ_{max} to two to three times λ_{min} or to 1.5 Å, whichever is higher.

The beam shutter must provide a fast and reproducible opening and closing. The older shutters of Laue beamlines at ESRF and Advanced Photon Source (APS) were able to operate in a few microseconds – a time that is comparable with the revolution time of a single particle bunch in the storage ring. To permit single X-ray pulse diffraction experiments, a fast shutter train has been devised (Bourgeois et al., 1996) that can isolate the individual X-ray pulse of around 100 ps duration, emitted by the particle bunch. Both circuitry and software have been developed and applied successfully (Bourgeois et al., 1996; Genick et al., 1997) that enable the shutter opening and closing to be linked to the master accelerator clock, and also to the triggering of an external device that initiates a structural change in the crystal, such as a pulsed laser. This is essential for rapid time-resolved Laue experiments.

The demands on the properties of a detector used for a Laue experiment are, in principle, the same as those for monochromatic data collection:

- a large dynamic range;
- a large active area both to record the reflections at higher resolution and to accommodate the larger crystal-to-detector distances desirable to minimize spatial overlaps of spots;
- a narrow point spread function so as not to impair the spatial overlap problem;
- a high detective quantum efficiency (DQE), especially for time-resolved experiments that generate weak diffraction patterns from the quite short exposures; and
- a fast readout that improves the Laue experiment duty cycle, since the readout time is always much longer than the very short Laue exposure times.

A fast readout is critical when an irreversible reaction is followed after a single reaction initiation, since the time resolution in this type of experiment is actually restricted by the detector readout time. Both image-plate and CCD detectors have been successfully used for Laue data collection, with their improved DQE allowing shorter exposure times and the automated readout changing the capacity to record many orientations for multiple Laue exposures compared to the film techniques used in early days of synchrotron Laue experiments.

As highlighted previously, several data sets at different crystal settings must be collected in order to obtain a final data set with the necessary accuracy and completeness. This is due to the fact that a single Laue image from a suitably oriented crystal will yield a substantial fraction of the unique data, but with low

redundancy and accuracy and poor coverage of reciprocal space at low resolution. Furthermore, the wavelength-normalization curve cannot be determined accurately from such a single exposure. However, the collection of several data sets at different crystal orientations overcomes these problems. A good Laue data set needs at minimum between five and 20 images, with the smaller number being appropriate to cases with high crystal symmetry.

The processing of Laue diffraction data requires special data evaluation packages. Efficient software packages were developed for this purpose, including LAUEVIEW (Ren and Moffat, 1995 a, b) and the Daresbury Laue Software Suite (Helliwell et al., 1989; Campbell, 1995). Improvements to specific parts of the processing were proposed: LEAP (Wakatsuki, 1993), LAUECELL (Ravelli et al., 1996; Ravelli, 1998), PrOW (Bourgeois et al., 1998) and other implementations based on Bayesian theory (Bourenkov et al., 1996; Ursby and Bourgeois, 1997). A flow chart of a typical Laue data processing is presented in Figure 8.4. As in the monochromatic data processing, this consists of three major parts: geometric prediction of diffraction patterns; integration of diffraction spots; and data reduction. Laue data reduction includes wavelength-normalization, frame-to-frame scaling, and harmonic deconvolution.

As mentioned above, the time-resolved crystallography requires use of the Laue method to allow data collection on the (typically rapid) time scales set by the reaction rates. Studies down to nanosecond time resolution are possible. These ultrafast studies need at least an equally rapid means of initiating the reaction, which translates into laser-induced photolysis or other photo-chemical reactions. In addition to being light-inducible, reactions to be studied on very rapid time scales should be reversible, as this allows the averaging of exposures (with the crystal having the same orientation) to improve the signal-to-noise ratio of the data and the collection of complete data sets from several crystal settings.

The photoactive yellow protein (PYP) is a very attractive system in which to study the molecular basis of signal transduction, as it undergoes a fully reversible, light-induced modification of the chromophore (Cusanovich and Meyer, 2003) and is therefore optimally suited to rapid, time-resolved crystallographic characterization. PYP is a bacterial photoreceptor which is believed to be involved in a negative phototactic response to blue light. After absorbing a blue light photon, the covalently attached coumaric acid chromophore undergoes *trans* to *cis* isomerization, which initiates a fully reversible photocycle that lasts on the order of 1 s. The photon energy is transduced into a structural signal as the molecule thermally relaxes through a series of spectroscopically distinguishable intermediates, in which the final two intermediates are denoted pP and pB. A time-resolved crystallographic Laue study on the E46Q mutant of PYP has been carried out (Anderson et al., 2004) with data collected after photoactivation at 30 time delays spaced evenly in the logarithm of time from 10 ns to 100 ms at the end of the photocycle. The photocycle was initiated by illumination with a 5- to 7-ns laser pulse. Experiments were performed near room temperature in order to avoid freezing out the pattern of transient structural change. The extent of photoactivation was unsatisfactory ($\sim$6% to 34%), and therefore data sets

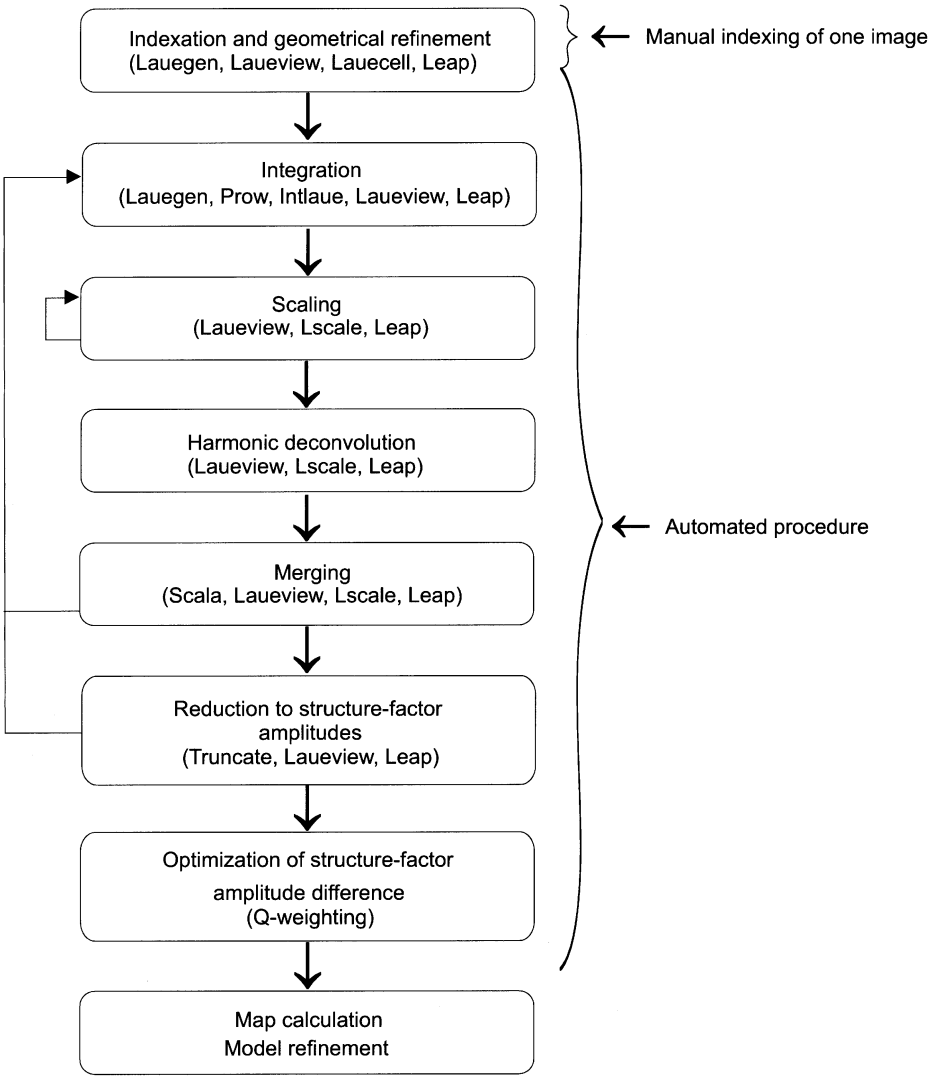


Fig. 8.4 A flow chart of typical Laue diffraction data processing.

were averaged in reciprocal space at adjacent time delays to increase the signal-to-noise ratio, effectively increasing data quality at the expense of a reduction in time resolution. The refinement of transient chromophore conformations shows that the spectroscopically distinct intermediates are formed via progressive disruption of the hydrogen bond network to the chromophore. Although structural change occurs within a few nanoseconds on and around the chromophore, it takes milliseconds for a distinct pattern of tertiary structural change to progress throughout the entire molecule, thus generating the putative signaling state.

Only two intermediate chromophore conformations are apparent from 10 ns to the end of the photocycle, and these have been denoted as pR and pB chromophore conformations. The pR conformation is the sole conformation present in the first three averaged electron density maps from 10 to 500 ns. The pB chromophore conformation is then present in the final four averaged electron density maps from 10 μs to 30 ms. The pR state is still close to the *trans* ground state, and the pB is the putative signaling state of PYP in the *cis* chromophore conformation. Thus, this time-resolved crystallographic study was able to structurally characterize two intermediate states of this rapid photochemical reaction.

Following a reaction on fast time scales is much more difficult if the reaction is nonreversible. There is a fundamental difference between systems that are inherently light-sensitive and those that must be rendered light-sensitive by chemical modifications. In the latter case, data must be collected at the same time points at different crystal settings in any case. The biggest problem is, how kinetically similar the crystals are at the respective data collection time points. Therefore, when studying fast, nonreversible reactions, it may be worth considering alternative approaches to time-resolved crystallography, such as trapping by low temperature. This also holds true if crystals are mosaic or become so during the reaction (initiation) since, as already mentioned, the Laue technique is extremely sensitive toward imperfections in the crystal. This was a severe problem in the time-resolved Laue experiments on the GTP complex of the Ras protein (Scheidig et al., 1994) and on ligand binding to glycogen phosphorylase b (Haijdu et al., 1987).

References

Anderson, S., Šrajer, V., Pahl, R., Rajagopal, S., Schotte, F., Anfinrud, P., Wulff, M., Moffat, K., *Structure* **2004**, *12*, 1039–1045.

Bolduc, J.M., Dyer, D.M., Scott, W.G., Singer, P., Sweet, R.M., Koshland, D.E., Stoddard, B.L., *Science* **1995**, *268*, 1312–1318.

Bourenkov, G.P., Popv, A.N., Bartunik, H.D., *Acta Crystallogr.* **1996**, *A52*, 797–811.

Bourgeois, D., Ursby, T., Wulff, M., Pradervand, C., LeGrand, A., Schildkamp, W., Laboure, S., Šrajer, V., Teng, T.-Y., Roth, M., Moffat, K., *J. Synchrotron Rad.* **1996**, *3*, 65–74.

Bourgeois, D., Nurizzo, D., Kahn, R., Cambillau, C., *J. Appl. Crystallogr.* **1998**, *31*, 22–35.

Campbell, J.W., *J. Appl. Crystallogr.* **1995**, *28*, 228–236.

Campbell, J.W., Habash, J., Helliwell, J.R., Moffat, K., *Int. Q. Protein Crystallogr.* **1986**, *18*, 23–31.

Cohen, B.E., Stoddard, B.L., Koshland, D.E., *Biochemistry* **1997**, *36*, 9035–9044.

Clifton, I.J., Duke, E.M.H., Wakatsuki, S., Ren, Z., *Methods Enzymol.* **1997**, *277*, 448–467.

Cusanovich, M.A., Meyer, T.E., *Biochemistry* **2003**, *42*, 965–970.

Duke, E.M., Wakatsuki, S., Hadfield, A., Johnson, L.N., *Protein Sci.* **1994**, *3*, 1178–1196.

Edman, K., Nollert, P., Royant, A., Beirhali, H., Pebay-Peyroula, E., Hajdu, J., Neutze, R., Laundau, E.M., *Nature* **1999**, *401*, 822–826.

Fülop, V., Phizackerley, R.P., Soltis, S.M., Clifton, I.J., Wakatsuki, S., Erman, J., Hajdu, J., Edwards, S.L., *Structure* **1994**, *2*, 201–208.

Genick, U., Borgstahl, G.E.O., Ng, K., Ren, Z., Pradervand, C., Burke, P.M., Šrajer, V., Teng, T.-Y., Schildkamp, W., McRee, D.E.,

Moffat, K., Getzoff, E.D., *Science* **1997**, *275*, 1471–1475.
Haijdu, J., Machin, P.A., Campbell, J.W., Greenhough, T.J., Clifton, I.J., Zurek, S., Gover, S., Johnson, L.N., Elder, M., *Nature* **1987**, *329*, 178–181.
Helliwell, J.R., Habash, J., Cruickshank, D.W.J., Harding, M.M., Greenhough, T.J., Campbell, J.W., Clifton, I.J., Elder, M., Machin, P.D., Papiz, M.Z., Zurek, S., *J. Appl. Crystallogr.* **1989**, *22*, 483–497.
Moffat, K., *Methods Enzymol.* **1997**, *277*, 433–447.
Moffat, K., *Chem. Rev.* **2001**, *101*, 1569–1581.
Petsko, G.A., *Methods Enzymol.* **1985**, *114*, 141–146.
Ravelli, R.B.G., *PhD thesis*, Utrecht University, Utrecht, **1998**.
Ravelli, R.B.G., Hezemans, A.M.F., Krabbendam, H., Kroon, J., *J. Appl. Crystallogr.* **1996**, *29*, 270–278.
Ren, Z., Moffat, K., *J. Appl. Crystallogr.* **1995a**, *28*, 461–481.
Ren, Z., Moffat, K., *J. Appl. Crystallogr.* **1995b**, *28*, 482–493.
Ren, Z., Bougeois, D., Helliwell, J.R., Moffat, K., Šrajer, V., Stoddard, B.L., *J. Synchrotron Rad.* **1999**, *6*, 891–917.
Rossi, G.L., *Curr. Opin. Struct. Biol.* **1992**, *2*, 816–820.
Scheidig, A.J., Sanchez-Llorente, A., Lantwein, A., Pai, E.F., Corrie, J.E.F., Reid, G.P., Wittinghofer, A., Goody, R.S., *Acta Crystallogr.* **1994**, *D50*, 512–520.
Schlichting, I., *Acc. Chem. Res.* **2000**, *33*, 532–538.
Schlichting, I., Goody, R.S., *Methods Enzymol.* **1997**, *277*, 467–490.
Schlichting, I., Almo, S.C., Rapp, G., Wilson, K., Petratos, K., Lentfer, A., Wittinghofer, A., Kabsch, W., Pai, E.F., Petsko, G.A., Goody, R.S., *Nature* **1990**, *345*, 309–315.
Schlichting, I., Berendzen, J., Phillips, G.N., Jr, Sweet, R.M., *Nature* **1994**, *371*, 808–812.
Schlichting, I., Berendzen, J., Chu, K., Stock, A.M., Maves, A.S., Benson, D.E., Sweet, R.M., Ringe, D., Petsko, G.A., Sligar, S.G., *Science* **2000**, *287*, 1615–1622.
Singer, P.T., Smålas, A., Carty, R.P., Mangel, W.F., Sweet, R.M., *Science* **1993**, *259*, 669–673.
Schmidt, M., Ihee, H., Pahl, R., Šrajer, V., *Methods Mol. Biol.* **2005**, *305*, 115–154.
Šrajer, V., Teng, T.Y., Ursby, T., Pradervand, C., Ren, Z., Adachi, S., Schildkamp, W., Bourgeois, D., Wulff, M., Moffat, K., *Science* **1996**, *274*, 1726–1729.
Stoddard, B.L., *Curr. Opin. Struct. Biology* **1998**, *8*, 612–618.
Stowell, M.H., McPhillips, T.M., Rees, D.C., Soltis, S.M., Abresch, E., Feher, G., *Science* **1997**, *276*, 812–818.
Ursby, T., Bourgeois, D., *Acta Crystallogr.* **1997**, *A53*, 564–575.
Verschueren, K.H.G., Seljée, F., Rozeboom, H.J., Kalk, K.H., Dijkstra, B.W., *Nature* **1993**, *363*, 693–698.
Wakatsuki, S., LEAP, Laue Evaluation Analysis Package, for time-resolved protein crystallography. In: Sawyer, L., Isaacs, N.W., Bailey, S. (Eds.), *Data Collection and Processing*, pp. 71–79. Daresbury Laboratory, Warrington, **1993**.
Yennewar, N.H., Yennewar, H.P., Farber, G.K., *Biochemistry* **1994**, *33*, 7326–7336.

9
Structural Genomics

9.1
Introduction

High-throughput DNA sequencing is now possible to determine the complete genomic sequences of whole organisms. Nowadays, the complete genomic sequences are available for many bacteria, a row of eukaryotic model systems including yeast, worm, the plants *Arabidopsis thaliana* and rice, fly, chicken, rat, mouse, monkey and last, but not least, for man. The genomic sequences have been annotated and can be publicly assessed at several data bases such as the EMBL Nucleotide Sequence Bank (Cochrane et al., 2006), which is produced in an international collaboration with GenBank (USA) and the DNA database of Japan (DDBJ). Each genomic "open reading frame" (ORF) codes potentially a protein with one or more biological functions. However, analysis of the genomes known so far indicates that a large fraction of the encoded proteins cannot be assigned to particular functions (or to particular pathways), and thus, no assays can be easily devised to investigate their exact roles. The existent recombinant techniques make it possible, in principle, to produce the gene product of each ORF of a genomic sequence in amounts for a functional or structural characterization. Since the function of a gene product is determined by its three-dimensional structure, this structure or its folding pattern may provide important insight into its biological function which, in turn, may help to place it in a particular cellular pathway. Efforts to define the three-dimensional structures of all gene products of a complete genome are known as structural genomics, and may provide an important foundation for the understanding of the biology of whole organisms.

A structural genomics approach involves:

- the selection of target proteins or domains;
- cloning, expression, purification and quality assessment of the targets;
- crystallization or labeling for nuclear magnetic resonance (NMR) spectroscopy;
- structure determination by X-ray crystallography or by NMR spectroscopy; and
- the archiving and annotation of new structures.

The structural genomics approach can be realized only if structures can be determined both quickly and cheaply, and this has focused attention on high-

throughput methods for protein production, characterization, crystallization, and structure determination. We have discussed such methods for crystallization, data collection at synchrotron beamlines and for structure determination in the respective previous chapters, and so will not refer to these again at this point. With the increasing success of these high-throughput methods, the structures of gene products might be defined in quite straightforward manner, perhaps within a few weeks. This will certainly be the case for bacterial targets, which can be easily produced in the bacterium *Escherichia coli*. Targets from eukaryotic organisms, however, can be prepared only with much more effort, and with a lower success rate than for their bacterial counterparts. The reasons for this will be discussed later in this chapter.

Structural genomics projects have been funded – and are currently being funded – in the United States, Canada, Japan and Europe, with a primary focus on proteins. However, a major secondary goal is to decrease the average cost through the development of high-throughput methods for all steps of structural genomics. Hence, in the following sections we will describe those components of structural genomics which have not been dealt with in previous chapters.

9.2 Target Selection

There are a very large number of sequences coding for proteins, particularly in eukaryotes, with for example approximately 30 000 genes in the human genome. Analyses of genome sequences have shown that the functions of very few proteins have previously been identified from genetics or biochemistry, although functions for about 50% can be deduced with reasonable confidence from knowledge of close homologs (e.g., Adams et al., 2000). With about 600 000 protein target sequences provided in the UNIPROT data base at the European Bioinformatics Institute (http://www.ebi.ac.uk), it is clear that a reasonable selection of targets is necessary. One approach is to select representatives of each homologous family, or even of each superfamily. This should be suited to determine all possible folds, which have been evolutionarily much more conserved than amino acid sequences, and should provide clues for the function of such a homologous family, assuming that the function can be deduced from the three-dimensional structure. The total number of homologous families can be operationally defined as the number of representative sequences whose neighbors (e.g., those within 30% amino acid sequence identity) jointly cover a certain percentage (e.g., 90%) of the sequence space. This number is probably 10 000 or more for a reasonable coverage (Linial and Yona, 2000; Vitkup et al., 2001). However, these homologous families can be grouped into a reduced number of superfamilies (about 1000 or so). One can assume that the great majority of the members of a homologous family will exhibit similar functions. For superfamilies that have divergently evolved, with conservation of general tertiary structure but perhaps less than 25% amino acid sequence identity, a related function may

be preserved. The selection of primary targets could be further focused by choosing representative structures of families only where there is no clue to function. Alternatively, one could choose only "core families" – those that are common to most genomes.

The identification of homologous families and superfamilies has been carried out using very sensitive approaches to sequence search and alignment. As an outcome of such studies, the Pfam data base has been created (Bateman et al., 2000), which currently holds about 8000 different protein domain families. Domains are the structural and functional building blocks of proteins, and so where the data are available, structural information has been used to ensure that Pfam families correspond to single structural domains. The domain boundaries used are currently those defined by the SCOP database (Murzin et al., 1995), and a new web-based tool allows direct cross-linking from domains on the SCOP web site to the corresponding Pfam families. This matching of families and domains enables an enhanced understanding of the function of multi-domain proteins.

Another approach is to select targets that are relevant to important biological functions, for example to human health. In this case, specific targets from man or pathogenic organisms must be selected, though this is very demanding as most of them are from eukaryotic organisms. Problems in preparing recombinant proteins from eukaryotes will be discussed in the next section.

However, a growing number of proteins have been found to be natively unfolded under physiological conditions (Wright and Dyson, 1999). Therefore, it is of key interest to predict from the amino acid sequence whether a given protein sequence is intrinsically unfolded, or not. Web-based tools can be used to perform this task. For example, the graphic web server, FoldIndex© (http://www.bioportal.weizmann.ac.il/fldbin/findex; Prilusky et al., 2005), is available. This is based solely on the average hydrophobicity of the protein's amino acids and the absolute value of its net charge, and has implemented the algorithm of Uversky et al. (2000).

9.3 Production of Recombinant Proteins

9.3.1 Introduction

The production of recombinant proteins in large amounts suitable for determination of their three-dimensional structure either by X-ray crystallography or NMR has been excellently explained in a contribution by Hughes and Stock (2001), and this has formed the basis for parts of this section. The production of recombinant proteins comprises several steps:

- to find an appropriate host for expression of the gene product;
- to design and produce a DNA segment that contains the gene coding for the desired gene product, which also contains all of the elements necessary for

high-level RNA expression and to be recognized by the host's translational machinery;
- to introduce the respective DNA segment into the host system;
- to grow the transformed hosts expressing the desired gene in amounts that will provide sufficient quantities of the expressed protein; and
- to isolate and purify the desired protein from the grown host material.

The recombinant protein, once expressed, needs to be folded correctly by the host or, if not, by the experimentalist.

The choice of host for expression depends on the nature of the gene product to be expressed. Bacterial protein will be properly expressed in bacterial hosts, and the bacterium *E. coli* is normally used as the standard expression system. Eukaryotic proteins are often post-translationally modified (cleavage, glycosylation, phosphorylation, etc.), and need such modifications in order to be properly expressed and folded. As a bacterial system is unable to perform these modifications, eukaryotic proteins must very often be expressed in eukaryotic hosts such as yeast, insect cells or mammalian cells. These individual points will be discussed in the following subsections.

9.3.2 Engineering an Appropriate Expression Construct

Initially, the most important decision to be made is which expression system should be used. As mentioned above, bacterial gene products can be properly expressed in prokaryotic (*E. coli*) expression systems, but for eukaryotic proteins the test of the *E. coli* expression system may also be useful. This is due to the ease of use of the *E. coli* expression system in terms of preparing the expression construct (cloning), growing the recombinant organism, and purifying the resulting protein. Furthermore, they allow for relatively easy incorporation of selenomethionine into the recombinant protein (Hendrickson et al., 1990), this being a prerequisite when applying the MAD technique to X-ray crystal structure determination. For many eukaryotic proteins, a prokaryotic expression system will be unsuitable for correct expression of the protein. However, this can be overcome by introducing changes in the construct that permit its expression in the prokaryotic system. If this fails, it will be necessary to move along the evolutionary pathway to yeast, insect cells, and finally to cultured mammalian cells. Although in these latter situations the problems associated with producing the protein in its native state are simpler, the difficulties of expressing large quantities of material both quickly and cheaply in an easy-to-purify manner become greater.

Once the expression system has been chosen, a suitable expression construct must be prepared. The first problem is to create the target gene, and this can be generated *de novo* from the respective genome sequence. However, this is a tedious task and a special field in its own right (e.g., Sambrook et al., 1989). Fortunately, cDNA clones or genomic DNA have been prepared for most of the

sequenced genomes, and can be acquired commercially. If only genomic DNA is available, it is easy to prepare the desired DNA clone using the polymerase chain reaction (PCR). Certain peculiarities must be taken into account if the targets of higher eukaryotes are to be expressed. Most of these genes contain introns, which are removed on the level of mRNA by a so-called splicing process. As *E. coli* and yeast have no (with minor exceptions) or different splicing machineries, respectively, cDNA clones must be used for the expression of such targets in these systems. Because some introns are large, cDNA clones are often used as basis of expression constructs in baculovirus systems, as well as in cultured insect and mammalian cells. From now on, for simplicity, we will assume that the cDNA of the target is available.

Optimizing the expression of the target protein is extremely important because it reduces the time and effort for growing the host cells or viruses, as well as simplifying the purification procedure. In order to generate the expression construct, DNA must be manipulated, and today this is done most effectively by using the PCR technique. For most constructs, the ends of the cDNA are modified by PCR with appropriate oligonucleotide primers that have been designed to introduce useful restriction sites and/or elements essential for efficient transcription and/or translation. Since it can often be advantageous to attempt the expression of a given target protein in a number of different vectors, it is useful to introduce carefully chosen restriction sites that enable the fragment either to be inserted simultaneously or to be transferred seamlessly into different plasmids or other vectors (Fig. 9.1). Figure 9.1 shows, schematically, the classical cloning technique with use of restriction and ligation enzymes. PCR can also be used to generate mutations within the cDNA. Since PCR may introduce mutations it is important to sequence all pieces generated by PCR after they have been cloned.

The addition of tags and domains, which are subsequently used for efficient purification by affinity chromatography, is now a standard technique and a prerequisite in structural genomics. These tags may be a small peptide or a larger protein, and can be added to the target sequence either at the amino or carboxyl terminus. The most common tags are hexahistidine (His_6), biotinylation peptides and streptavidin-binding peptides (Strep-tag), glutathione S-transferase (GST) and maltose-binding protein (MBP). The list of possible tags is much greater, and grows permanently. The introduction of such affinity tags creates a new problem: whether or not to remove the fused element. There are examples where leaving the tags did not perturb the crystallization for larger fusion elements such as GST and MBP, but in many cases these additional elements hindered the crystallization. Thus, it is advisable to remove these tags following the affinity chromatography step. For this purpose, a cleavage site for a specific protease must be incorporated between the additional element and the target protein.

In structural genomics target sequences will need to be cloned multiple times into many different expression vectors. However, the cloning of large sets of targets by conventional methods is impractical, for numerous technical reasons. Most importantly, these methods involve target-specific restriction analyses and

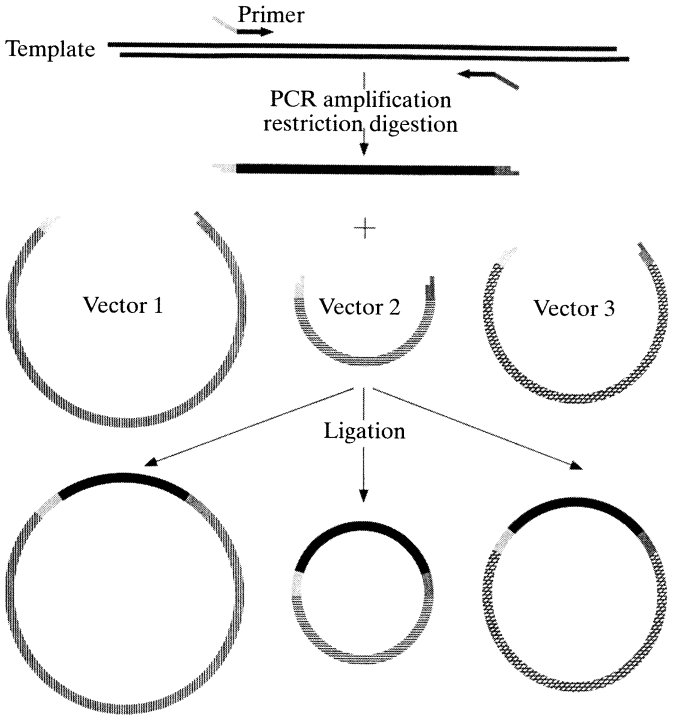


Fig. 9.1 The classical technique for creating an expression construct. PCR can be used to amplify the target sequence. PCR primers should be constructed to contain one or more restriction sites that can be used conveniently to subclone the fragment into the desired expression vector. It is often possible to choose vectors and primers such that a single PCR product can be ligated to different vectors. (Reproduced by permission of International Union of Crystallography, from Hughes and Stock, 2001.)

rely heavily on the purification of DNA fragments from agarose gels. Altogether, conventional methods cannot be automated and thus are not compatible with high-throughput projects. Several alternative cloning systems based on different recombination reactions have been described. Here, we briefly describe the GATEWAY® recombinatorial cloning system (Walhout et al., 2000), which has been used frequently in structural genomics projects (e.g., Vincentelli et al., 2003). GATEWAY is based on the recombination reactions that mediate the integration and excision of phage λ into and from the *E. coli* genome, respectively (Fig. 9.2 a). The integration involves recombination of the attP site of the phage DNA with the attB site located in the bacterial genome. This generates an integrated phage genome flanked by attL and aaR sites. The integration reaction (Fig. 9.2 a) needs two enzymes: the phage protein integrase (Int), and the bacterial protein integration factor (IHF) referred to as "BP Clonase". The recombination reaction is reversible; thus, the phage DNA can be excised from the bacterial genome by recombination between the attL and attR sites (Fig. 9.2 a, "LR re-

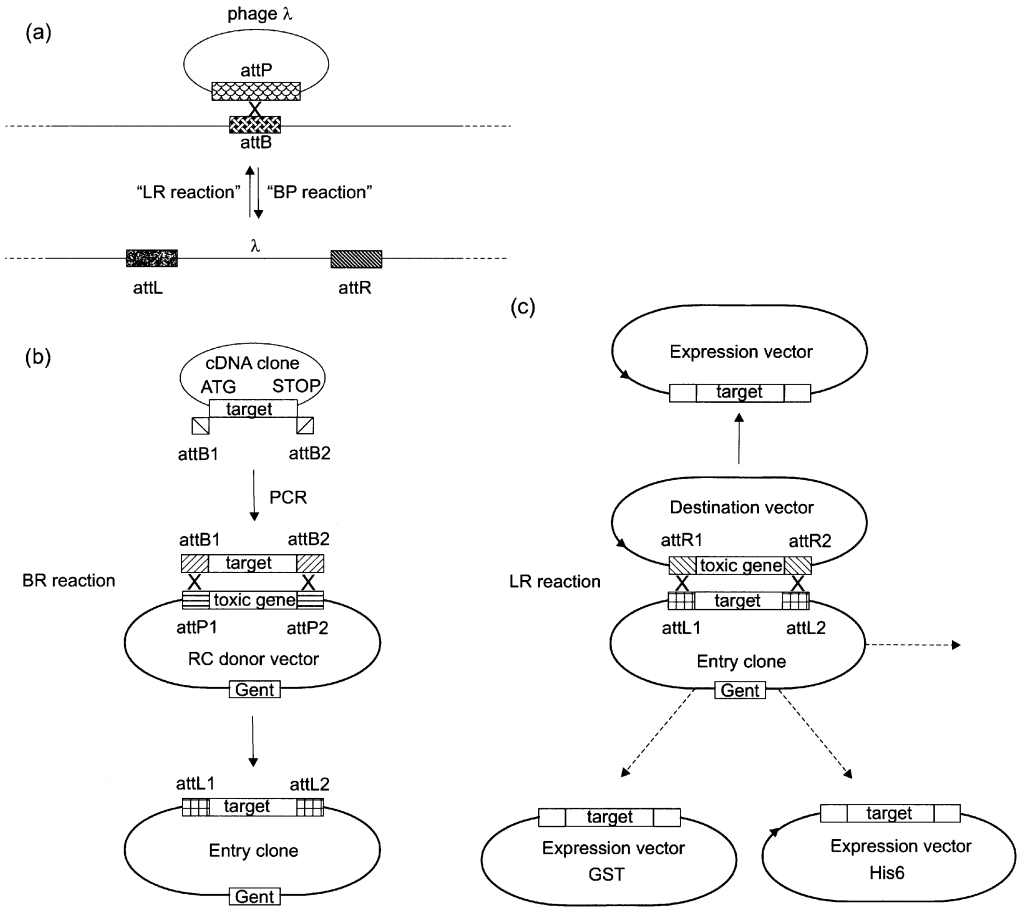


Fig. 9.2 The cloning procedure in the GATEWAY system. (a) Recombinational cloning (RC) is based on the integration of phage λ into the *E. coli* genome. (b) Insertion of the target gene into the RC donor vector by BR reaction to yield the entry clone. (c) Transfer of the target gene from entry clone by LR reaction into a given destination clone to yield the respective expression vector. (Adapted from Walhout et al., 2000.)

action"). This reaction requires Int, IHF and an additional phage enzyme, excionase (Xis) (collectively referred to as "LR Clonase").

BR and LR clonases have been purified, allowing the GATEWAY reactions to take place *in vitro*. Furthermore, the att sites have been mutated to generate pairs of derivatives in such a way that the corresponding B and L sites can recombine only with the relevant P and R sites, respectively. The duplication of att sites permits two independent recombination reactions to take place in the same molecules, one at the 5′ end of the target to be cloned and the other at the 3′ end.

The GATEWAY procedure takes the following route. PCR products corresponding to the target sequence flanked by attB1 and attB2 sequences are cloned into a

recombinational cloning (RC) donor vector (Fig. 9.2b). This donor vector holds a toxic gene flanked by attP1 and attP2 sequences and an antibiotic resistance marker, in this case for gentamicin. The result is that entry clone now has the inserted target sequence flanked by attL1 and attL2 sequences. Once a target sequence is cloned into an entry vector it can be transferred by RC reaction into different destination vectors that contain the toxic gene flanked by attR1 and attR2 sites (Fig. 9.2c). In Figure 9.2c the RC reaction is shown for a transfer into a general destination vector, and the expression vectors for His_6 and GST tags are displayed. The transfer into other destination vectors is possible, of course.

As result of the LR reaction, the target sequences are flanked by the 25-bp attB1 and attB2 sites in the resulting expression clones. This leads to an extra eight amino acids at both N- and C-terminal ends of the expressed proteins, but this should not influence the protein's behavior in terms of folding and subsequent crystallization. Nevertheless, this point should be borne in mind when using the GATEWAY system, in order to avoid unsatisfactory results.

9.3.3 Expression Systems

9.3.3.1 *E. coli*

In case the protein does not have extensive post-translational modifications, it is usually appropriate to begin with an *E. coli* host-vector system (for an extensive review of expression in *E. coli*, see Makrides, 1996). Both plasmid and viral-based (M13, λ, etc.) expression systems are available for *E. coli*. Although viral-based vector systems are quite useful for some purposes (expression cloning of cDNA strands, for example), in general, for the expression of relatively large amounts of recombinant protein, they are not as convenient as plasmid-based expression systems. Although there are minor differences in the use of viral expression systems and plasmid-based systems, the rules that govern the design of the modified segment are the same. Therefore, the more frequently used plasmid-based systems will be discussed, ranging from design of plasmid to fermentation conditions.

A large number of different, easy-to-use expression plasmids for *E. coli* are available (for a concise review, see Unger, 1997). In most cases, it is possible to identify expression and/or fermentation conditions that result in the production of a recombinant protein in amounts of greater than 5 mg L^{-1} of culture, which makes the scale of fermentation reasonable.

E. coli systems are either constitutive (they always express the coded protein) or inducible, where a specific change in the culture conditions is necessary to induce the expression of the recombinant protein. If the desired protein is toxic to *E. coli*, then an inducible system must be used. The induction may be created by a temperature shift, as in systems using the bacteriophage λ p_L promoter and the temperature-sensitive repressor CI857ts, or by the addition of an inducer such as isopropyl-β-D-thiogalactopyranoside (IPTG). IPTG-induced systems can be under the control of the *lac* repressor (e.g., $lacI^q$) or a *lac*-controlled operon that encodes the bacteriophage T7 RNA polymerase (e.g., Studier et al.,

1990). In this system, the respective operon is contained in the genome of the *E. coli* host with one copy in the cell only. Induction with IPTG leads to the synthesis of the T7 RNA polymerase, which recognizes a promoter sequence that is different from the sequence recognized by *E. coli* RNA polymerase. If such an *E. coli* system also carries a multicopy plasmid, in which the target sequence is linked to a T7 promoter, the T7 RNA polymerase efficiently produces mRNA from the plasmid. This usually leads to the production of a large amount of the desired recombinant protein. *E. coli* strains that carry a *lac*-inducible T7 RNA polymerase are readily available, as are the respective expression plasmids that hold the T7 promoters. In *E. coli* the initiation of translation requires not only an appropriate initiation codon (usually AUG, occasionally GUG), but also a special element, the Shine-Dalgarno sequence, just 5′ of the initiator AUG. In *E. coli*, the first step in translation involves binding of the 30S ribosomal subunit and the initiator fMET-tRNA to the mRNA. The Shine-Dalgarno sequence is complementary to the 3′ end of the 16S RNA found in the 30S subunit. Eukaryotic mRNAs do not contain Shine-Dalgarno sequences. Some *E. coli* expression plasmids carry a Shine-Dalgarno sequence, others do not. If one is not present in the plasmid, it must be introduced when the cDNA sequences is modified before introduction into the expression plasmid.

Problems during expression may be caused by proteolytic cleavage by cellular proteases, most notably ClpA, at the N-terminus with N-terminal amino acids Phe, Leu, Trp, Tyr, Arg or Lys. Furthermore, codon usage may influence expression levels. Unfortunately, there are substantial differences in preferences/usage in prokaryotes and eukaryotes. The expression of eukaryotes in *E. coli* may be improved by optimizing the codon usage of the eukaryotic target sequence to be expressed.

Fermentation is an especially important part of protein expression. By using an identical strain or plasmid, slight alterations in growth conditions can make substantial differences in the yield of the protein. In this situation there are many parameters to be optimized, including the media, the temperature of fermentation and, in a larger fermenter, the aeration and stirring. Very often it is necessary to develop new fermentation conditions when scaling up the fermentation; indeed, this is a particular problem when the scale is changed from shake flasks to a fermenter.

When screening expression constructs for production of recombinant protein, four scenarios are most commonly encountered:

- high-level expression of soluble recombinant protein;
- high-level expression of the recombinant protein with a greater or lesser proportion of the protein in inclusion bodies;
- no expression, or very low expression;
- lysis of cell.

The first of these outcomes is usually the most welcome. It may happen that not all soluble protein molecules are properly folded, with misfolded proteins occasionally being expressed at high levels in soluble form. Such proteins usual-

ly exhibit aberrant behavior during purification, such as precipitation, migration as broad peaks during column chromatography, and elution in the void volume during size-exclusion chromatography. In such cases, additional experimentation is required. Inclusion bodies are usually the result of improper protein folding, and cell lysis generally indicates severe toxicity. There are two obvious reasons for failure to produce measurable amounts of recombinant protein: (i) there is a problem at the level of transcription and/or translation; or (ii) the protein is being degraded proteolytically.

At this point we will not discuss potential solutions to these problems in detail (for more information, see Hughes and Stock, 2001), but two issues deserve mention. The formation of inclusion bodies is the result of aggregation of unfolded protein molecules. The protein molecules may not interact properly with the *E. coli* chaperones, or the chaperones are overstrained by the high concentration of recombinant protein in the host cell. The expression of the protein into inclusion bodies has both positive and negative consequences. Proteins in inclusion bodies are essentially immune to proteolytic degradation. Additionally, it is usually relatively easy to obtain the inclusion bodies in relatively pure form, making it simple to purify the recombinant protein. A variety of protocols are available for refolding (e.g., De Bernadez Clark, 1998), but there are few simple, universal, procedures.

Proteolytic degradation is an active process in *E. coli*, and several strategies for minimizing the proteolysis of recombinant proteins have been developed (Enfors, 1992; Murby et al., 1996). These strategies include the secretion of proteins into the periplasm or external media, the engineering of proteins to remove proteolytic cleavage sites, and growth at low temperatures, as well as other strategies to promote folding, such as the use of fusion proteins and coexpression with chaperones.

9.3.3.2 **Eukaryotic Expression Systems**

9.3.3.2.1 **Yeasts**

Yeasts are simple eukaryotic cells. Indeed, considerable effort has been expended in studying brewers' yeast, *Saccharomyces cerevisiae*, and in developing plasmid systems and expression systems that can be used in such systems. Recently, methylotrophic yeasts – most notably *Pichia pastoris* – have been developed as alternative systems that offer several advantages over *S. cerevisiae*. Although yeast systems are reasonably robust, the expertise required to use them effectively is less frequently available than the respective expertise for the manipulation of *E. coli* strains. Nor are the tools, media and reagents necessary to grow yeast and select for the presence of the expression plasmids broadly available as those used for *E. coli* systems. However, the increasing commercial availability of complete kits (such as *Pichia* expression systems from Invitrogen) is making yeast systems more accessible. While yeast systems do offer some advantages relative to *E. coli*, these advantages are, in general, modest. Specifically,

the problem of mimicking the post-translational modifications found in higher eukaryotes (particularly glycosylation), which has not been solved for *E. coli*, has not yet been solved in yeast either. None of the available systems recapitulates the post-translational modifications found in higher eukaryotes.

9.3.3.2.2 **Baculovirus**

Baculovirus expression systems are becoming increasingly important tools for the production of recombinant proteins for X-ray crystallography. The insect cell-virus expression systems are more experimentally demanding than bacterial or yeast, but they offer several advantages. Because insects are higher eukaryotes, many of the difficulties associated with the expression of proteins from higher eukaryotes in *E. coli* do not occur. There is no need for a Shine-Dalgarno sequence, there are no major problems with codon usage, and fewer problems occur with a lack of appropriate chaperones. Although glycosylation is not the same in insect and mammalian cells, in some cases it is close enough to be acceptable. Baculovirus systems allow expression at reasonable levels, typically ranging from 1 to 500 mg L^{-1} of cell culture. Considerable effort has put into the development of convenient transfer vectors, and today baculovirus expression kits are available from more than ten commercial sources.

Baculoviruses usually infect insect cells; in terms of the expression of foreign proteins, the important baculoviruses are the *Autographa californica* nuclear polyhedrosis virus (AcNPV) and the *Bombyx mori* nuclear polyhedrosis virus (BmNPV). Baculovirus expression vectors have been reviewed widely (e.g., Jones and Morikawa, 1996; Merrington et al., 1997; Possee, 1997). It is interesting here briefly to describe the mode of operation of the baculovirus expression system in insect cells. In nature, in the late stage of replication in insect larvae, nuclear polyhedrosis viruses produce an occluded form, in which the virions are encapsulated in a crystalline protein matrix, termed polyhedron. When the virus is released from the insect larvae, this proteinaceous coat protects the virus from the environment and is necessary for its propagation in the natural state. However, replication of the virus in cell cultures does not require the formation of occlusion bodies. In tissue culture, the production of occlusion bodies, is dispensable and the primary protein, polyhedron, is not required for replication. Cultured cells infected with wild-type AcNPV produce large amounts of polyhedron, and the same applies to cells infected with modified AcNPV vectors, with other genes inserted in place of the polyhedron gene (or in place of another highly expressed gene, *p10*, that is dispensable in cultured cells).

There are several important points to consider when setting up the baculovirus cell-culture system. Although most baculoviruses have a relatively restricted host range and AcNPV was first isolated from alfalfa looper (*Autographa californica*), for the purpose of expressing foreign proteins, it is usually grown in cells of the armyworm (*Spodoptera frugiperda*) or cabbage looper (*Trichoplusia ni*). The isolation and purification of the appropriate AcPNV vectors are normally carried out in monolayer cultures, whereas large quantities of recombinant protein are normally pro-

duced in suspension cultures. There is also the issue of whether fetal calf serum should be included in the culture medium, or not.

Compared to bacteria and yeast cells, cells from higher eukaryotes are quite delicate, and considerable care must be taken in cell culture. The cells are subject to shear stress, which can be a problem in stirred and/or shaken cultures. Compared to bacterial and yeast cells, cultured cells grow relatively slowly and require rich media that will support the rapid growth of a wide variety of unwanted organisms; hence, special care must be taken in order to avoid contaminating the cultures. Although antibiotics are commonly used for this purpose, they will not in general prevent contaminations with yeasts or molds, which often cause the greatest problems. Thus, a very neat mode of operation is needed.

9.3.3.2.3 **Mammalian Cells**

In some cases, however, the baculovirus and/or insect cell expression systems are unable to make the desired recombinant protein product, but if the target is sufficiently important it can be produced in cultured mammalian cells. It must be noted, however, that the effort required to produce tissue cultured cells which express high levels of recombinant protein is substantial. Cell lines are usually prepared by transfection, after which some of the cells will incorporate transfected DNA into their genomes. A number of agents can be used to transfect DNA; these include (but are not limited to) calcium phosphate, DEAE dextran, and cationic lipids. This is a complex and poorly defined process, and the transfected DNA is often incorporated into complex tandem arrays. Neither the amount of transfected DNA nor the location in the host genome is controlled in a standard transfection, and consequently the expression level varies substantially from one transfected cell to another. This makes the process of creating mammalian cell lines that efficiently and stably express a recombinant protein very labor-intensive. Other transfection methods include electroporation or homologous recombination.

As mentioned above, tissue culture cells are much more difficult to grow than those of either yeast or *E. coli*, and this of course is also valid for mammalian cell cultures. Proteins for successful X-ray structure determinations have been prepared, for example, from cultures of Chinese hamster ovary (CHO) cells or immortalized human B cell clones. Human embryonic kidney (HEK) 293 cells have been recently tested in the structural genomics project SPINE (Aricescu et al., 2006) with promising results. These are adhering cells which are relatively robust, easy to culture, and have a good growth rate.

9.3.4 **Protein Purification**

In the past, when proteins had to be purified from natural materials, purification ratios of up to 5000 were not unusual. However, since the development of efficient systems to express recombinant protein, combined with protein constructs provided with tags for affinity chromatography, the required purification

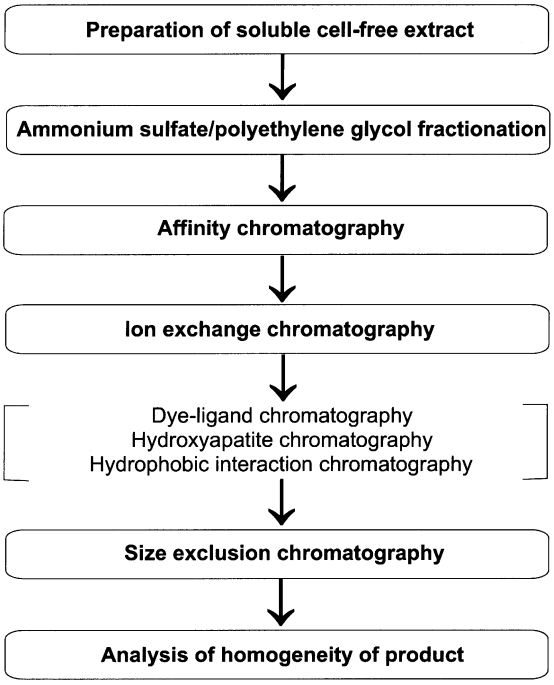


Fig. 9.3 The general scheme for protein purification.

ratios have been reduced to about 20–50. In particular, this improvement has also notably diminished the amount of basic raw material required. Furthermore, purification systems have been automated and several types of high-capacity, high-flow rate chromatography media and columns have been developed and are now commercially available.

If reasonably good levels of expression can be achieved, then most recombinant proteins can be purified using a relatively simple procedure, as depicted in Figure 9.3. All purification steps are based on the fact that the biochemical properties of proteins differ: proteins have different sizes, surface charges, and hydrophobicities.

9.3.4.1 **Precipitation**

Precipitation is often used as the first step in a purification procedure, and in part it can be used to separate proteins from nucleic acids. Nucleic acids are highly charged polyanions, and their presence in a protein extract can dramatically reduce the efficiency of column chromatography, for example by the saturation of anion-exchange resins. The most commonly used precipitation reagents are ammonium sulfate and polyethylene glycols (PEGs). If the precipitation range is broad, it is very efficient simply to precipitate the majority of pro-

teins by adding ammonium sulfate to 85% saturation, or 30% PEG 6000. Precipitation can also serve as a useful method for concentrating proteins at various steps during purification, as well as for storing proteins that are unstable upon freezing or upon storage in solution.

9.3.4.2 **Chromatography**

Column chromatography steps in which the protein is adsorbed onto the resin under one set of conditions and then eluted from the column under a different set of conditions, can result in significant protein purification. In case the protein construct holds an affinity tag, the first chromatographic step will be an affinity chromatography. Because the process of affinity chromatography is so powerful (purification ratios of up to 100 can be achieved) and the development of a specific affinity column is difficult, general procedures for affinity chromatography have been developed. In Section 9.3.2 we have already mentioned the most common tags, namely His_6, biotinylated peptides and Strep-tag, GST and MBP. The column material for such an affinity chromatography is produced by linking the respective substrate or another binding entity to an inert support. The desired protein binds selectively to the column, and can usually be eluted by washing the column with the same substrate used to prepare the column (e.g., glutathione for GST-tagged protein bound to a glutathione agarose column) or with a competing metal ligand such as imidazole for a His_6-tagged protein bound to a nickel-nitrilotriacetic acid (Ni-NTA) column.

In a purification scheme without affinity tag, anion-exchange chromatography is usually a good starting point. Most proteins have acidic pIs, and conditions can often be found that allow binding of the protein to anion-exchange matrices. If conditions cannot be found under which the protein binds to an anion-exchange resin, then binding to a cation-exchange column can be attempted. Fewer proteins interact with cation-exchange resins, but if the desired protein does bind this can be a powerful step. The use of an anion-exchange column does not necessarily preclude the use of a cation-exchange column; under appropriately chosen sets of conditions (most notably adjustment of pH), a given protein can bind to both resins. Dye–ligand, hydroxyapatite (a variation of ion-exchange chromatography) and hydrophobic interaction chromatographies may represent beneficial improvements in the purification procedure.

Size-exclusion chromatography (SEC), which does not involve adsorption of the protein onto the matrix, rarely provides as much purification as the chromatography steps described above. However, this can be a good step to include at the end of a purification scheme. The isolation of a well-defined peak in the included volume separates intact, properly folded protein from any damaged/aggregated species that may have been generated during the purification procedure. Although SEC does not provide a definitive analysis of such behavior, migration of the protein consistent with its expected molecular weight is a good sign. The elution of a relatively small protein in the void volume suggests a need for further analysis. The described purification steps may also be very beneficial for affinity-tagged pro-

tein following the affinity chromatography step. It may also be possible to introduce an affinity tag at both the N- and C-termini; in this case, the respective affinity chromatographies are performed one after the other.

With proteins fused to an affinity tag, there is also the issue of whether to remove the tag, or not. Tag removal usually involves engineering a site for a specific protease, digestion with the protease, and subsequent purification to isolate the final cleaved product.

As mentioned above, expressing high levels of recombinant prokaryotic and eukaryotic proteins in *E. coli* can lead to the production of improperly folded material that aggregates into insoluble inclusion bodies. Following lysis of the cells, inclusion bodies can usually be recovered relatively easily by low-speed centrifugation (5 min at 12 000 *g*). Inclusion bodies are larger than most macromolecular structures found in *E. coli*, and more dense than *E. coli* membranes. In most cases, the inclusion bodies contain the desired recombinant protein in relatively pure form, whereupon the problem lies not with the purification of the protein but rather in finding the correct way to refold it. Various procedures for refolding proteins from inclusion bodies have been described (e.g., De Bernadez Clark, 1998). The insoluble inclusion bodies are usually solubilized in a powerful chaotropic agent such as guanidine hydrochloride or urea, after which the denaturant is sequentially removed by dilution, dialysis, and/or filtration. After refolding, the properly folded soluble protein must be separated from the fraction that did not fold appropriately. In this respect, incorrectly folded proteins are relatively insoluble and can generally be removed by centrifugation. Once the soluble protein has been obtained, conventional purification procedures may be employed and the integrity of the purified protein should be checked.

9.3.5
Quality Control of the Purified Protein

Before starting with crystallization screenings, the properties and purity of the recombinant protein should be carefully checked. Although several proteins crystallize well from relatively impure preparations, it is advisable to use highly purified proteins for crystallization trials. There are several reasons for this. First, it is easier to obtain the high concentrations of protein ($>10\ mg\ mL^{-1}$) normally needed for crystallization if the protein is pure and the behavior of highly purified protein is more reproducible. A homogeneous preparation of protein will precipitate at a specific point, rather than over a broad range of solution conditions. Furthermore, degradation during storage and/or crystallization is minimized if all of the proteases have been removed.

The most convenient and widely used methods to check protein purity involve electrophoresis, SDS-PAGE, and/or isoelectric focusing. SDS-PAGE may be slightly more convenient for the detection of unrelated proteins: isoelectric focusing is probably more useful in detecting subspecies of the recombinant protein target.

If the preparation is relatively free of unrelated protein, but there is concern about the presence of multiple species of the desired recombinant protein target, then several techniques can be applied. Mass spectroscopy is capable of detecting small differences in molecular weights, and for proteins up to several hundred amino acids in length it is generally able to detect differences in mass equivalent to a single amino acid. This can be useful in detecting heterogeneity in post-translational modifications (if present), and in detecting heterogeneity at both the N- and C-termini.

In terms of crystallization, the ability to produce a highly concentrated monodisperse protein preparation is probably more important than absolute purity. A number of methods can be used to determine whether or not the protein is aggregating. SEC has been widely used, mainly by biochemists, but in structural genomics projects dynamic light scattering has been used routinely to check concentrated protein preparations for aggregation (Vincentelli et al., 2003). The method is relatively simple, it is very sensitive to small amounts of aggregation, and it has the additional advantage that the protein sample can be used for subsequent crystallization trials, because it is not consumed by the method. If simple heterogeneity is detected, one is faced with the problem of whether this will adversely affect the crystallization and, if so, how to remove it. The alteration of the expression construct may provide improved results.

Usually, the protein will have been produced in larger amounts and will not be totally consumed in the crystallization experiments. Thus, the produced protein must be properly stored. As a general rule, it is better to store proteins as highly purified solutions (concentrations >1 mg mL^{-1}). If the protein contains oxidizable sulfurs, reducing agents can be added and the solution held in a non-reducing (N_2) atmosphere. It is essential that the protein be stored in a manner that will not allow microbial growth, which is normally achieved by sterilization of the protein solution by filtration through 0.2 μm filters and/or the addition of antimicrobial agents such as NaN_3. For long-term storage (periods longer than a few weeks), protein solutions are often precipitated in ammonium sulfate or frozen either at –20 or –80 °C; for this they should preferably be divided into aliquots in order to avoid repeated thawing of the protein. Freezing samples at intermediate concentrations (1–3 mg mL^{-1}) is usually more effective than freezing either extremely dilute or concentrated samples. Cryoprotective agents can be added to protein samples destined to be frozen, but it must be borne in mind that the cryoprotectant may not be desired in the crystallization experiment. Thus, after thawing the sample, the cryoprotectant must be removed.

9.4 Aspects of Automation

The need for high-throughput methods in structural genomics projects was discussed in Section 9.1, the automation of crystallization in Section 1.5, and of structure determination in Sections 2.3.2 and 6.5, as well as in Chapter 5. To-

day, the automation of protein production and characterization has been established in all structural genomics projects in different variants, depending on the expression system used. The main objective of automation is to process multiple samples in parallel, saving both time and costs, as well as generating consistent and reproducible results. The process of automating a system constitutes two distinct parts: miniaturization and automation. In general, automated systems perform liquid-handling tasks in multi-well plates, with preferred volumes of 1 mL, or less.

The highest degree of automation can be achieved in the *E. coli* expression system. A typical automated high-throughput protein production system based on *E. coli* has been developed at the Joint Center for Structural Genomics (McMullan et al., 2005) and will be described here briefly. In order to maximize flexibility and minimize costs, a conventional cloning approach was chosen. Subsequently, a robotic platform was developed, which incorporates both liquid and plate handling, with thermocyclers and a plate reader. In this way a work flow of up to 384 validated expression clones per week could be achieved. To allow expression at a scale sufficient for crystallization trials, the group developed a parallel fermentation system (GNFermenter), for parallel 96-culture high-density growth that produces 2–4 g of cell pellet. Processing of the resulting cell pellets through affinity purification is performed with custom automation (GNFuge). The fermentation tubes are directly processed in the GNFuge, for the steps of lysis, removal of cell debris, and affinity purification, after which the resulting purified proteins can be processed by secondary purification or advanced directly to quality assessment and crystallization.

A more miniaturized system based on *E. coli* with N-terminal His_6-tag has been developed by Finley et al. (2004). The system was designed to process up to 384 unique samples in parallel, using the GATEWAY cloning and expression system in four 96-well plates. All liquid handling steps are carried out in 96-well plates in volumes of 1 mL or less and, with the exception of the bacterial plating and colony picking, all steps have been completely automated, including pipetting robots and PCR plate incubators.

The automation of protein production in eukaryotic expression systems is much more complicated, and preliminary attempts and solution for such systems have been discussed by Aricescu et al. (2006).

References

Adams, M.D., Celniker, S.E., Holt, R.A., et al., *Science* **2000**, *287*, 2185–2195.

Aricescu, A.R., Assenberg, R., Busso, D., Chang, V.C., Davis, S.J., Dubrovsky, A., Gustafsson, L., Hedfalk, K., Heinemann, U., Jones, I.M., Ksiazek, D., Lang, C., Maskos, K., Messerschmidt, A., Macieira, S., Peleg, Y., Parrakis, A., Poterszman, A., Schneider, G., Sixma, T., Sussman, J.L., Sutton, G., Tarboureich, N., Zeev-Ben-Mordehai, T., Jones, E.Y., *Acta Crystallogr.* **2006**, *D62*, 1114–1124.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., Sonnhammer, E., *Nucleic Acids Res.* **2000**, *28*, 263–266.

Cochrane, G., Adelbert, P., Althorpe, N., Anderson, M., Baker, W., Baldwin, A., Apweiler, R., et al., *Nucleic Acids Res.* **2006**, *34*, D10–D15.
De Bernadez Clark, E., *Curr. Opin. Biotechnol.* **1998**, *9*, 157–163.
Enfors, S.-O., *Trends Biotechnol.* **1992**, *10*, 310–315.
Finley, J.B., Qiu, S.-H., Luan, C.-H., Luo, M., *Prot. Expr. Purif.* **2004**, *34*, 49–55.
Hendrickson, W.A., Horton, J.R., LaMaster, D.M., *EMBO J.* **1990**, *9*, 1665–1672.
Hughes, S.H., Stock, A.M., Preparing recombinant proteins for X-ray crystallography. In: Rossmann, M.G., Arnold, E. (Eds.), *International Tables for Crystallography*, Vol. F, pp. 65–80, Kluwer Academic Publishers, Dordrecht, **2001**.
Jones, I., Morikawa, Y., *Curr. Opin. Biotechnol.* **1996**, *7*, 512–516.
Linial, M., Yona, G., *Prog. Biophys. Mol. Biol.* **2000**, *73*, 297–320.
Makrides, S.C., *Microbiol. Rev.* **1996**, *60*, 512–538.
McMullan, D., Canaves, J.M., Quijano, K., Abdubek, P., Nogoghossian, E., Haugen, J., Klock, H.E., Vincent, J., Hale, J., Paulsen, J., Lesley, S.A., *J. Struct. Funct. Genomics* **2005**, *6*, 135–141.
Merrington, C.L., Bailey, M.J., Possee, R.D., *Mol. Biotechnol.* **1997**, *8*, 283–297.
Murby, M., Uhlén, M., Ståhl, S., *Protein Exp. Purif.* **1996**, *7*, 129–136.
Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C. *J. Mol. Biol.* **1995**, *247*, 536–540.
Possee, R.D., *Curr. Opin. Biotechnol.* **1997**, *7*, 569–572.
Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I., Sussman, J.L., *Bioinformatics* **2005**, *21*, 3435–3438.
Sambrook, J., Fritsch, E.F., Maniatis, T., *Molecular Cloning: A Laboratory Manual*, 2nd Ed., Cold Spring Harbor Laboratory Press, New York, **1989**.
Studier, F.W., Rosenberg, A.H., Dunn, J.J., Dubendorff, J.W., *Methods Enzymol.* **1990**, *185*, 60–89.
Unger, T.F., *The Scientist* **1997**, *11*, 20–23.
Uversky, V.N., Gillespie, J.R., Fink, A.L., *Proteins: Struct. Funct. Genetics* **2000**, *41*, 415–427.
Vincentelli, R., Bignon, C., Guez, A., Canaan, S., Sulzenbacher, G., Tegoni, M., Campanacci, V., Cambillau, C., *Acc. Chem. Res.* **2003**, *36*, 165–172.
Vitkup, D., Melamud, E., Moult, J., Sander, C., *Nat. Struct. Biol.* **2001**, *8*, 559–566.
Walhout, A.J., Temple, G.F., Brasch, M.A., Hartley, J.J., Lorson, M.A., van den Heuvel, S., Vidal, M., *Methods Enzymol.* **2000**, *382*, 575–592.
Wright, P.E., Dyson, H.J., *J. Mol. Biol.* **1999**, *293*, 321–331.

Part II
Practical Examples

Introductory Remarks

This part of the book is not intended to be a tutorial for an X-ray crystal structure analysis. In fact, to write such a book would be completely impossible because of the many different steps involved in these methods. However, there are further points of view which must be taken into account. The ideal case would be that a researcher could pursue the determination of a 3D structure, from selection of the biomacromolecule through production, crystallization and X-ray structure analysis itself to annotation and functional characterization. Whilst this entire procedure would be difficult to achieve for one person, a growing number of researchers in the field have a biological background and wish to determine for themselves the 3D structure of their favorite biomacromolecule. Consequently, a series of explanations utilizing practical examples should be beneficial not only for this group of readers but also – hopefully – for those with a wider chemical or physical background.

In Part 1 of this book we discussed the practical aspects of protein production, crystallization and X-ray data collection; thus, a more detailed argument would extend the scope of the book. Hence, Part 2 contains a collection of examples for key steps in an X-ray crystal structure determination, starting with a suitable X-ray diffraction data set, which may comprise several individual data sets from different crystals, heavy-atom derivatives, or collected at different wavelengths.

10
Data Evaluation

10.1
Autoindexing, Refinement of Cell Parameters, and Reflection Integration

One prerequisite for a successful and reliable X-ray crystal structure determination of a biological macromolecule is sufficiently well-diffracting crystals. If such crystals are available, the decision must then be made as to which radiation source and experimental set-up is most appropriate for an optimal diffraction data collection. When the quality of the crystals is very good, data collection using in-house X-ray equipment may be sufficient, but in most circumstances it is much more advantageous to carry out the data collection at a synchrotron beamline. For molecular replacement and high-resolution projects, a beamline with fixed wavelengths is adequate, but a wavelength-tunable beamline is necessary for the application of MAD techniques, including a heavy-atom derivative crystal suitable for MAD. A microfocus beamline is indispensable for all circumstances when the crystal is small. The crystals are measured at cryogenic temperatures in almost all cases, apart from those situations where the crystal cannot be frozen.

Here, the data evaluation will be explained by means of the crystal structure determination of the enzyme 4-hydroxy-butyryl-CoA dehydratase (4-BUDH) from *Clostridium aminobutyricum* (Martins et al., 2004). We will use this example to illustrate other steps of crystal structure determination that follow.

The enzyme 4-BUDH catalyzes the reversible oxygen-sensitive dehydration of 4-hydroxybutyryl-CoA and the oxygen-insensitive isomerization of vinyl-CoA to crotonyl-CoA. It is active as homotetramer with up to one $[4Fe\text{-}4S]^{2+}$ cluster and one noncovalently bound flavin adenine dinucleotide (FAD) moiety per 54-kDa subunit. As the enzyme contains a $[4Fe\text{-}4S]^{2+}$ cluster, the MAD technique could be applied to solve the phase problem. The respective diffraction data collection was carried out at the synchrotron beamline PX at the Swiss Light Source at the Paul Scherrer Institute, Villigen, Switzerland, at cryogenic temperatures. Data sets were gathered at three different wavelengths: (i) remote ($\lambda = 0.90004$ Å); (ii) at the Fe peak ($\lambda = 1.73652$ Å); and (iii) at the Fe inflection point ($\lambda = 1.74314$ Å). We will discuss the diffraction data evaluation on the basis of the remote data set. The program MOSFLM has been chosen to demonstrate course of action.

The MARCCD detector at the PX beamline produced a set of diffraction images of type *img*, which can be input directly to the MOSFLM program. MOSFLM must be installed on your computer, together with the CCP4 suite. We use the Linux operating system in all of our practical demonstrations. The following actions must be taken in order to run MOSFLM:

1. Copy the whole set of *img*-files into a respective directory on your computer.
2. Create an input file similar to that in Figure 10.1. All rows starting with a ! are treated as comments. One has to assign a TITLE, a directory for the GENERATE-File, a directory for the image-files, an identifier of the image-files, and the scanner type. You may supply the crystal detector distance DIST and the beam center individually, but this is also automatically read from the image file(s). Furthermore, one must provide the actual wavelength, resolution, mosaic spread and first image file to be read. Cell, matrix and symmetry have been commented out because we have no prior information about the unit cell and symmetry of the crystal.
3. Enter "ipmosflm" from the Linux prompt.

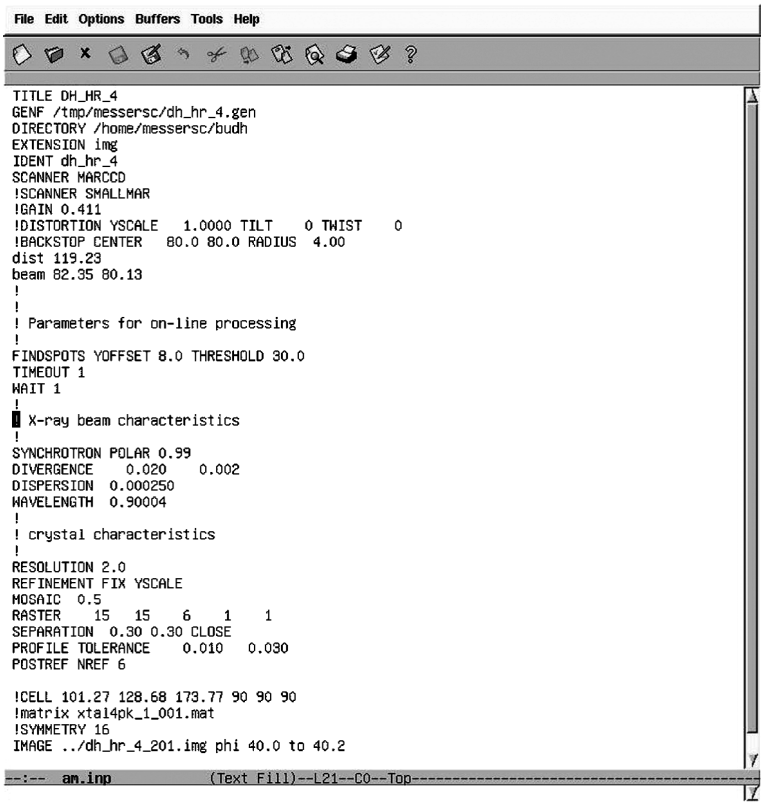


Fig. 10.1 The initial input file for MOSFLM for running in graphical mode.

4. Enter "@start.inp" from the MOSFLM prompt, here "@am.inp". The graphical user interface of MOSFLM appears (Fig. 10.2), and the left window displays the parameters we have input so far. The right window shows the actual diffraction pattern with well-resolved diffractions spots and an acceptable mosaicity.
5. Click the "Find spots" button. This generates a list of reflections, which serves as input for the Autoindexing procedure. One can use several images for the Autoindexing. If desired, you must read in the respective images and find the spots.
6. Click the "Autoindex" button. One must confirm some default settings, after which the results of the autoindexing appear (Fig. 10.3). The penalty is a measure of the deviation of the symmetry-corrected unit cell with the observed unit cell. One chooses the combination of lowest penalty and highest symmetry. In our example this is No. 4. We select space group P222 from the four possibilities, because at this stage we do not know if systematic extinction are present, which would indicate the existence of twofold screw axes. All four putative space groups belong to the same Laue group, and we will prove the

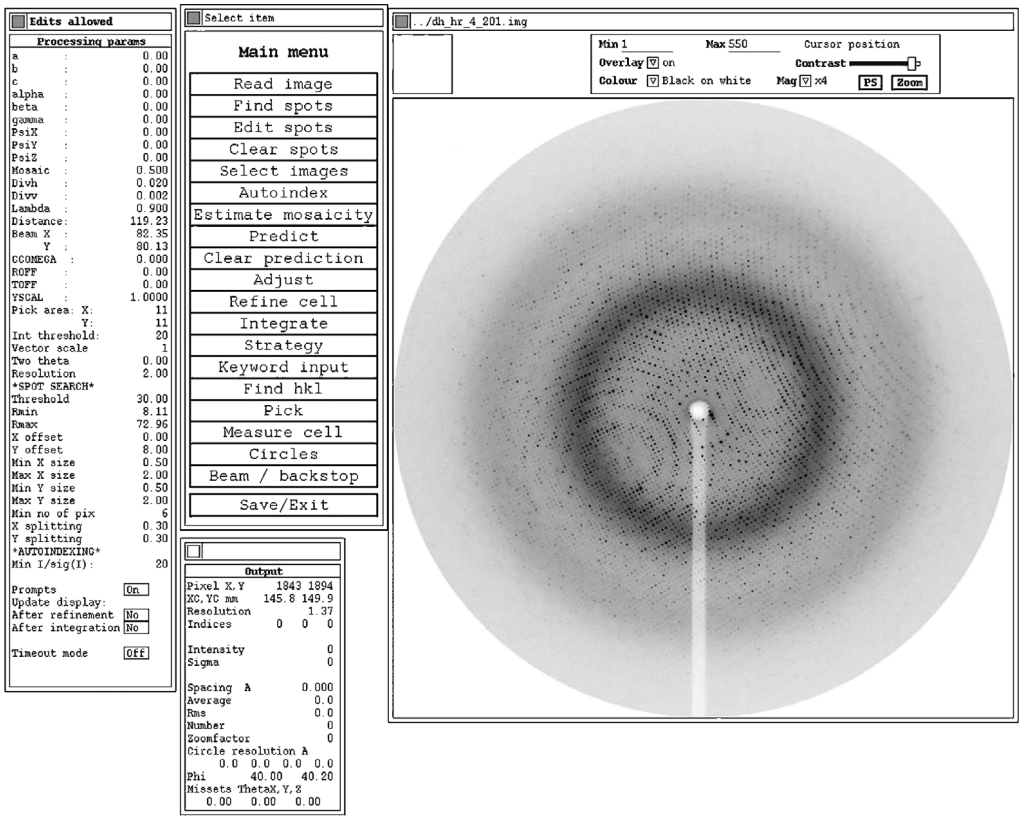


Fig. 10.2 Graphical user interface of MOSFLM after input of start file as displayed in Figure 10.1.

Input reply

```
List of possible Laue groups, sorted on penalty index.
The lower the PENALTY, the better
Only solutions with PENALTY less than 200 are listed, a complete list is given in the terminal window
 No PENALTY  LATT    a        b        c     alpha  beta gamma  Possible spacegroups
 16 169      mC   101.10   361.13   128.69    90.0  90.2  73.8   C2
 15 169      oC   101.10   361.13   128.69    90.0  90.2 106.2   C222,C2221
 14 169      mC   361.13   101.10   128.69    90.2  90.0  73.8   C2
 13 168      mC   101.10   276.26   173.43    90.0  90.1  68.7   C2
 12 168      mC   276.26   101.10   173.43    90.1  90.0  68.7   C2
 11 168      oC   101.10   276.26   173.43    90.0  90.1 111.3   C222,C2221
 10 106      tP   101.10   128.69   173.43    90.0  90.1  90.2   P4,P41,P42,P43,P422,P4212,P4122,P41212,P4222,P42212,P4322,P43212
  9 105      mC   163.89   163.43   173.43    90.1  90.1 103.7   C2
  8 105      oC   163.43   163.89   173.43    89.9  90.1  76.3   C222,C2221
  7 105      mC   163.43   163.89   173.43    89.9  90.1  76.3   C2
  6   1      mP   101.10   173.43   128.69    90.0  90.2  90.1   P2,P21
  5   1      mP   128.69   101.10   173.43    90.1  90.0  90.2   P2,P21
  4   1      oP   101.10   128.69   173.43    90.0  90.1  90.2   P222,P2221,P21212,P212121
  3   1      mP   101.10   128.69   173.43    90.0  90.1  90.2   P2,P21
  2   0      aP   101.10   128.69   173.43    90.0  89.9  89.8   P1
  1   0      aP   101.10   128.69   173.43    90.0  90.1  90.2   P1

Suggested Solution:  4  P222
           penalty:  1
              cell: 101.105  128.691  173.433  90.00  90.10  90.17
  regularized cell: 101.105  128.691  173.433  90.00  90.00  90.00
          Symmetry: oP (Primitive Orthorhombic)


Select a solution AND a spacegroup from list above (eg 3 p42) or 0 to abandon or T to change min I/sig(I):4 P222␣
```

Fig. 10.3 Output window of Autoindexing for 4-BUDH. Solution number 4 with lowest penalty (triclinic solutions have a penalty of 0, *per se*) and highest symmetry (primitive orthorhombic compared to primitive monoclinic) has been chosen.

correctness of this by calculating the R_{merge} value later. The result of the Autoindexing was very promising and one can move to the unit cell refinement as the next step.

7. Click the "Refine cell" button. There appears an input window (Fig. 10.4). In this example, images from three different segments are used with three images per segment. The refined orientation matrix is stored in the file "dh_hr_2_201_ref.mat". The starting spindle axis values and increments are also indicated. The unit cell refinement is now running automatically and, if successful, the results will be displayed (Fig. 10.5). The refinement worked smoothly, with little changes in the unit cell parameters, detector distance and missetting angles. Large deviations are suspicious and may be an indication that either something went wrong in previous steps or that the data set cannot be evaluated at all.

At the end of the unit cell refinement the graphical user interface has the appearance of Figure 10.6. All actual parameters are displayed. The coincidence of the predicted and observed spots can be checked by pressing the "Predict" button, whereupon the predicted spots appear as colored boxes. If there is no per-

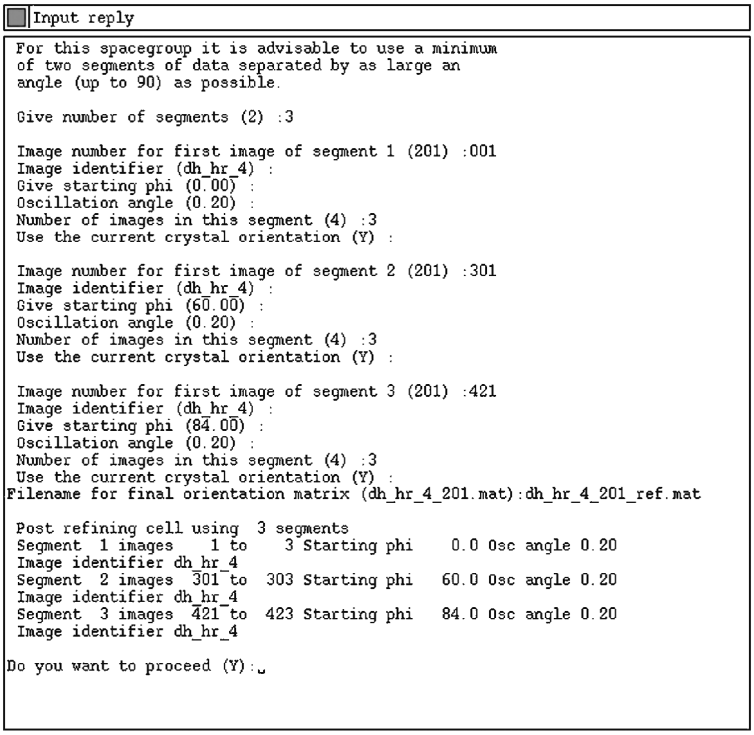


Fig. 10.4 Input window for the refinement of the unit cell and orientation parameters.

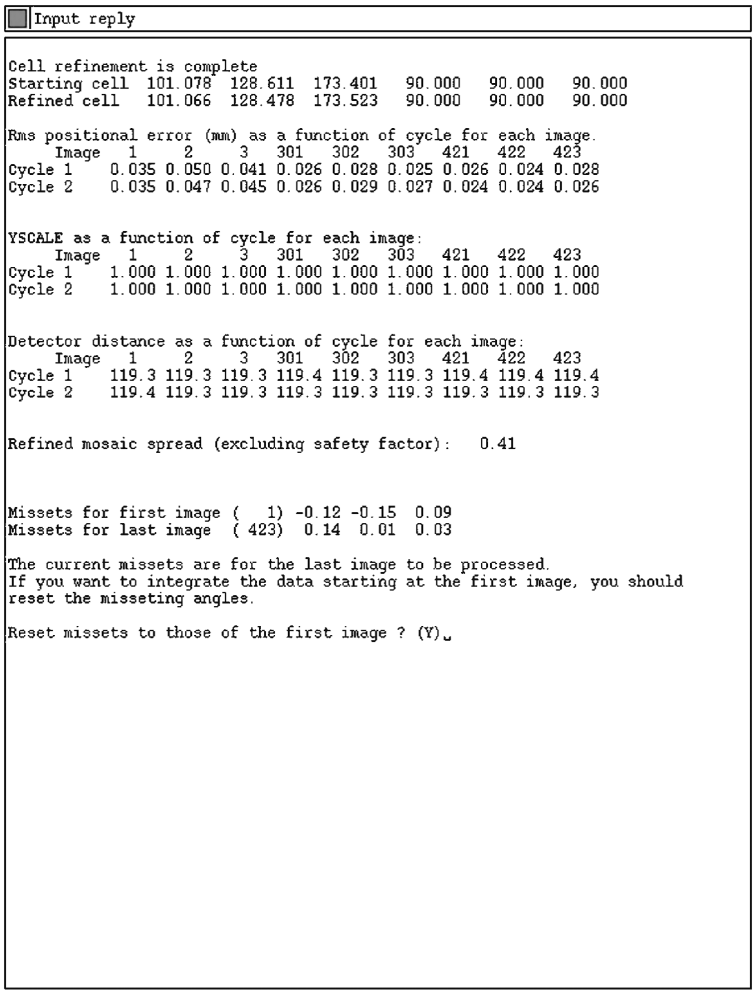


Fig. 10.5 Results of the "Refine cell" step for 4-BUDH.

fect match something has failed in the previous steps. One can now start the reflection integration, which is best done in batch mode. One exits the graphical user interface and the actual parameters are saved in a file on request.

Figure 10.7 shows the input file for running MOSFLM in batch mode. The actual parameters can be copied from the save file produced when exiting the Graphical User Interface (GUI) of MOSFLM. The refined orientation matrix is contained in file "dh_hr_4_201_ref.mat" and is invoked by the keyword MATRIX. Due to the small spindle axis increment of 0.2° per image, most of the reflections are partials. Therefore, MOSFLM instructions "REFINE INCLUDE PARTIALS" and "PROFILE PARTIALS" must be included. It is advisable to fix

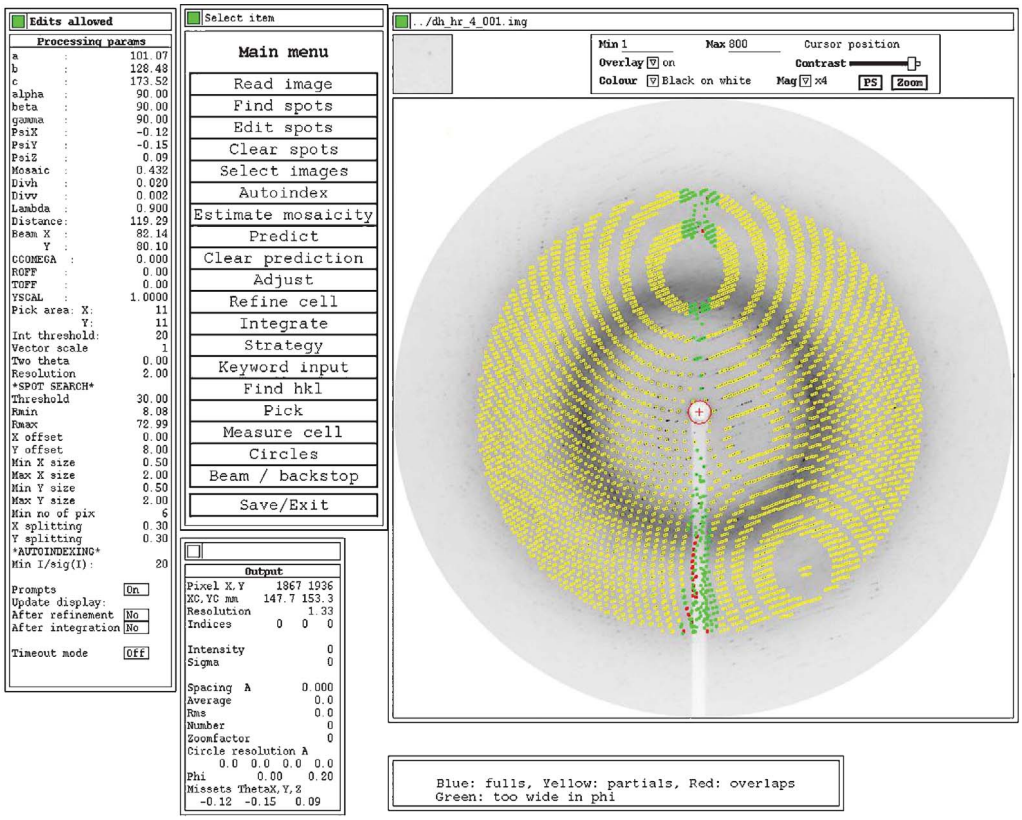
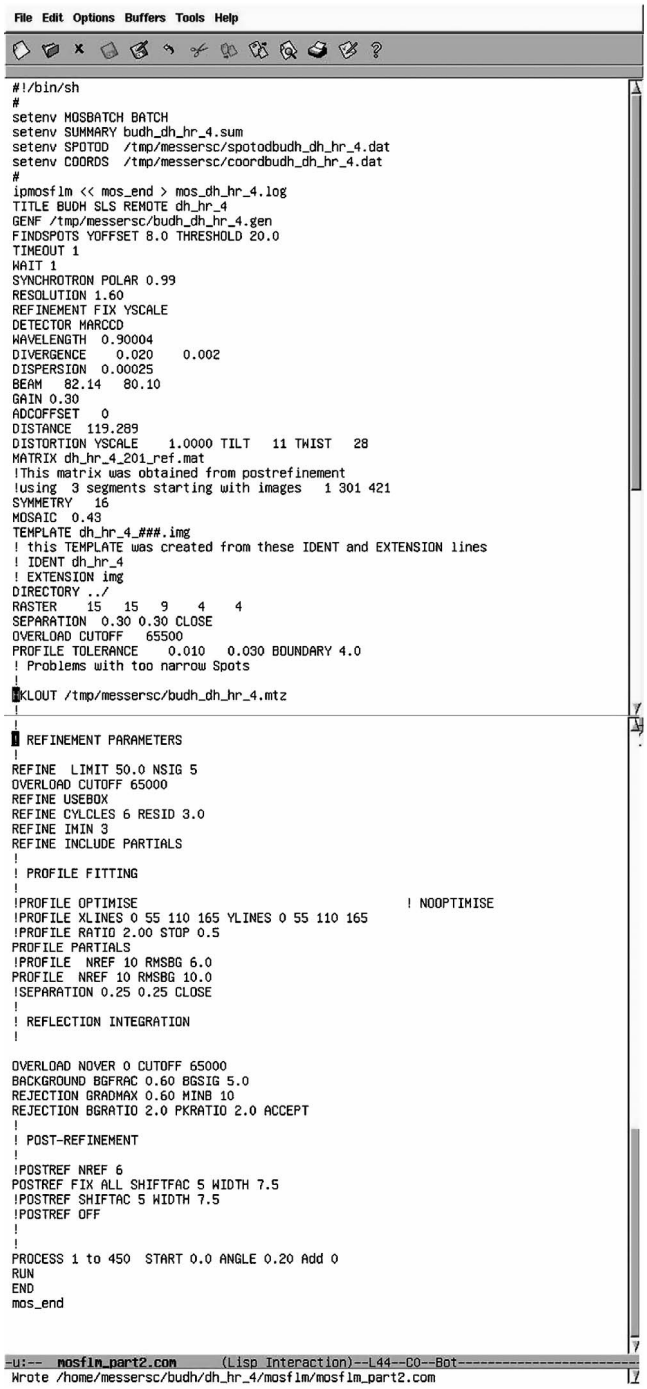


Fig. 10.6 GUI of MOSFLM for 4-BUDH after the "Refine cell" step. The actual parameters are displayed in the left window. The predicted spots positions are shown.

the unit cell parameters, which have been refined before, during the post-refinement step. This is done by specifying "POSTREF FIX ALL". Finally, the reflection integration is done for all 450 images of the data set (PROCESS 1 to 450 START 0.0 ANGLE 0.20). Image 1 starts at spindle angle 0.0° and the angle increment is 0.2° per image. This job produces a large log file (here, "mos_dh_hr_4.log"), a summary file (here, "budh_dh_hr_4.sum") and the reflection output file in CCP4 mtz-file style (here, "budh_dh_hr_4.mtz"). The essential information of the log file is contained in the summary file. At the end of this file there is a table displaying the refined detector parameters and R_{sym} values for all individual images. As in many cases of data collection the crystal is in an arbitrary orientation, the individual images contain few or no symmetry related reflections. If no symmetry reflections were present, the R_{sym} value is given as 0.0, or if only a few reflections have contributed its value is not meaningful. The last table lists the results of the post-refinement. As the unit cell pa-



```
File  Edit  Options  Buffers  Tools  Help
#!/bin/sh
#
setenv MOSBATCH BATCH
setenv SUMMARY budh_dh_hr_4.sum
setenv SPOTOD  /tmp/messersc/spotodbudh_dh_hr_4.dat
setenv COORDS  /tmp/messersc/coordbudh_dh_hr_4.dat
#
ipmosflm << mos_end > mos_dh_hr_4.log
TITLE BUDH SLS REMOTE dh_hr_4
GENF /tmp/messersc/budh_dh_hr_4.gen
FINDSPOTS YOFFSET 8.0 THRESHOLD 20.0
TIMEOUT 1
WAIT 1
SYNCHROTRON POLAR 0.99
RESOLUTION 1.60
REFINEMENT FIX YSCALE
DETECTOR MARCCD
WAVELENGTH  0.90004
DIVERGENCE    0.020    0.002
DISPERSION  0.00025
BEAM   82.14   80.10
GAIN 0.30
ADCOFFSET   0
DISTANCE  119.289
DISTORTION YSCALE    1.0000 TILT   11 TWIST   28
MATRIX dh_hr_4_201_ref.mat
!This matrix was obtained from postrefinement
!using  3 segments starting with images   1 301 421
SYMMETRY   16
MOSAIC  0.43
TEMPLATE dh_hr_4_###.img
! this TEMPLATE was created from these IDENT and EXTENSION lines
! IDENT dh_hr_4
! EXTENSION img
DIRECTORY ../
RASTER    15   15   9    4    4
SEPARATION  0.30 0.30 CLOSE
OVERLOAD CUTOFF   65500
PROFILE TOLERANCE    0.010   0.030 BOUNDARY 4.0
! Problems with too narrow Spots
!
BKLOUT /tmp/messersc/budh_dh_hr_4.mtz
!
!
! REFINEMENT PARAMETERS
!
REFINE  LIMIT 50.0 NSIG 5
OVERLOAD CUTOFF 65000
REFINE USEBOX
REFINE CYLCLES 6 RESID 3.0
REFINE IMIN 3
REFINE INCLUDE PARTIALS
!
! PROFILE FITTING
!
!PROFILE OPTIMISE                                    ! NOOPTIMISE
!PROFILE XLINES 0 55 110 165 YLINES 0 55 110 165
!PROFILE RATIO 2.00 STOP 0.5
PROFILE PARTIALS
!PROFILE  NREF 10 RMSBG 6.0
PROFILE  NREF 10 RMSBG 10.0
!SEPARATION 0.25 0.25 CLOSE
!
! REFLECTION INTEGRATION
!

OVERLOAD NOVER 0 CUTOFF 65000
BACKGROUND BGFRAC 0.60 BGSIG 5.0
REJECTION GRADMAX 0.60 MINB 10
REJECTION BGRATIO 2.0 PKRATIO 2.0 ACCEPT
!
! POST-REFINEMENT
!
!POSTREF NREF 6
POSTREF FIX ALL SHIFTFAC 5 WIDTH 7.5
!POSTREF SHIFTAC 5 WIDTH 7.5
!POSTREF OFF
!
!
PROCESS 1 to 450  START 0.0 ANGLE 0.20 Add 0
RUN
END
mos_end

-u:--  mosflm_part2.com      (Lisp Interaction)--L44--C0--Bot----------------------
Wrote /home/messersc/budh/dh_hr_4/mosflm/mosflm_part2.com
```

Fig. 10.7 Input file for running MOSFLM in batch mode for 4-BUDH.

rameters were fixed, the only relevant parameters are the missetting angles PHIX, PHIY, PHIZ and the effective mosaic spread. In this particular example, the maximum shift was less than the 0.25° that is acceptable. The effective mosaic spread was refined to 0.43°, a value low enough to deliver a high-quality intensity data set.

10.2 Scaling of Intensity Diffraction Data

At this stage, the intensities have been integrated individually in each diffraction image. As highlighted in Section 4.6, they must be scaled and averaged. The different data evaluation systems such as MOSFLM (SCALA), HKL2000 (SCALEPACK) or XDS (XSCALE) use their own scaling programs (given in parentheses). The mathematics behind these programs is quite similar and has been described in Section 4.6. In our MOSFLM example, we use the CCP4 program SCALA. The relevant input file is listed in Figure 10.8. First, the data on our MOSFLM reflection output file ("budh_dh_hr_4.mtz") must be sorted, as in the CCP4 routine "sortmtz". The sorted file is input to routine "scala". As these data have been collected in one batch we have one run command only. However, it is possible that one could not evaluate the data set in one MOSFLM run due to strong crystal slippage, or the data had to be collected from several crystals due to considerable radiation damage. In this case, one would have several MOSFLM output reflections files to be sorted in "sortmtz", and take into account the different batches by relevant run instructions. Intensities must be integrated from partials, and anomalous scattering must be switched on because we want to use information from anomalous diffraction. The "scales batch" command has many options. The one used here is recommended for anomalous data and integration of partials.

The last part is a CCP4 "truncate step", which reduces all reflections to the unique reflections while keeping track of the anomalous information and individual reflections contributing to the unique reflection. The CCP4 mtz reflection file has a record for each reflection holding the indices H, K, L, the intensity I with standard deviation SIGI, or structure factor amplitude F with standard deviation SIGF and a row of other putative items. This data set has been collected at remote wavelength. The anomalous difference DANO plus its standard deviation has been stored to allow each data set to be treated in a general manner later in the program SHARP, which will be used for the phase calculation and refinement. Except for H, K, L, the other items can be given individual labels, which has not be done in this example (e.g., F=F, which could also be F=<*desired label*>). The truncated reflection file is output to "hklout" (here, "budh_dh_hr_trn.mtz").

The job generates a longer log-file, with two relevant tables, which are also needed for the data submission to the Protein Data Bank or for a scientific publication. The first one is the data evaluation analysis against resolution

```
File  Edit  Options  Buffers  Tools  Help
#!/bin/sh
#
# first need to sort mtz file output from mosflm or rotaprep
#
#goto SCALA ! csh-specific
sortmtz hklout \
            /tmp/messersc/budh_dh_hr_4_sort.mtz <<EOF-sort
H K L M/ISYM BATCH I SIGI
/tmp/messersc/budh_dh_hr_4.mtz
EOF-sort
#
#SCALA: ! csh-specific
#
# Scala - calculating batch scale factors & merging
# Simple case - single scale and B factor for each batch
# see $CEXAM/unix/non-runnable/scala.exam for other examples
# /fs/scratch/scr
scala hklin /tmp/messersc/budh_dh_hr_4_sort.mtz \
      hklout   /tmp/messersc/budh_dh_hr_4_mrg.mtz \
      scales   /tmp/messersc/budh_dh_hr_4.scales \
      rogues   /tmp/messersc/budh_dh_hr_4.rogues \
      normplot /tmp/messersc/budh_dh_hr_4.norm \
      anomplot /tmp/messersc/budh_dh_hr_4.anom \
       << eof-scala
run 1 batch 1 to 450
sdcorr  1.25 0.03
resolution 58.0 1.60
partials maxwith 10
exclude SDMIN 3.00
REJECT SCALE 6.00 6.00 REJECT
REJECT MERGE 6.00 6.00 KEEP
scales batch brotation spacing 5.0 tails
## Alternative simple scaling models:
#  1) batch scales, smooth Bfactor (recommended for synchrotron data)
# scales batch brotaion spacing 5
#  2) smooth scaling (recommend for laboratory data or "dose-mode" collection)
# scales rotation spacing 5
##
intensities integrated partials
anomalous on
final partials
eof-scala
#
#
truncate hklin /tmp/messersc/budh_dh_hr_4_mrg.mtz \
   hklout budh_dh_hr_4_trn.mtz <<EOF-trunc
title BUDH REMOTE dh_dr_4 SLS  Data a - mosflm 1.60 Angstrom 450 images
nresidue 6400
labout F=F SIGF=SIGF DANO=DANO SIGDANO=SIGDANO
#labout  F=FP SIGF=SIGFP
EOF-trunc
-u:--  budh_scala_input.com      (Lisp Interaction)--L1--C0--Top----------
```

Fig. 10.8 The SCALA input file for 4-BUDH.

(Fig. 10.9). Rmgr is the R_{merge} value (Eq. 4.26), and its cumulative value should not exceed values of about 0.15 as a rule of thumb. For this 4-BUDH data set it is 0.055, which is very satisfying. I/sigma(I) is the ratio of the mean intensity to the mean standard deviation of the intensity, and should be not less than about 2 for the last resolution bin (here, a value of 1.8 was still accepted). Figure 10.10 displays the completeness and multiplicity against resolution. The completeness should be close to 100%, which is the case in this example. The overall multiplicity is 3.6, which is a satisfying value.

Until now, we had evaluated assuming the space group P222. The scaling had proved that the assumption of the orthorhombic Laue group is correct due

File Edit Options Buffers Tools Help

N	1/d^2	Dmin(A)	Rmrg	Rfull	Rcum	Ranom	Nanom	Av_I	SIGMA	I/sigma	sd	Mn(I)/sd	Nmeas	Nref	Ncent	FRCBIAS	Nbias
$$																	
1	0.0391	5.06	0.026	0.000	0.026	0.020	7817	26804.	1174.9	22.8	2012.	23.8	29241	8447	1144	0.014	12973
2	0.0781	3.58	0.027	0.000	0.027	0.016	14850	40312.	1703.1	23.7	3080.	23.7	51976	15012	1118	0.026	24097
3	0.1172	2.92	0.034	0.000	0.029	0.020	20522	22385.	1175.4	19.0	1794.	21.4	71059	20729	1206	0.029	33207
4	0.1562	2.53	0.046	0.000	0.032	0.029	24684	10779.	748.5	14.4	962.	18.2	84345	24866	1239	0.016	39650
5	0.1953	2.26	0.061	0.000	0.035	0.039	28163	7203.	670.9	10.7	742.	15.5	95281	28271	1247	0.000	45247
6	0.2344	2.07	0.082	0.000	0.039	0.057	31181	5149.	598.4	8.6	659.	12.6	104044	31273	1263	-0.005	49525
7	0.2734	1.91	0.115	0.000	0.043	0.087	34027	3185.	500.2	6.4	567.	9.3	112451	34131	1282	-0.003	54337
8	0.3125	1.79	0.176	0.000	0.047	0.136	36555	1732.	410.6	4.2	477.	6.3	117162	36405	1255	-0.008	58571
9	0.3516	1.69	0.268	0.000	0.051	0.207	38673	1066.	393.5	2.7	448.	4.2	121165	38092	1183	-0.002	61349
10	0.3906	1.60	0.405	0.000	0.055	0.314	40742	712.	393.4	1.8	452.	2.9	125911	39834	1163	0.003	63746
$$																	

```
">&lt;b&gt;For inline graphs use a Java browser&lt;/b&gt;&lt;/applet&gt;

 Overall:        0.055  0.000  0.055  0.039 277214    8087.   742.3  10.9   874. 11.3  912635 277060 12100   0.016 442702
                  Rmrg  Rfull   Rcum  Ranom  Nanom    Av_I   SIGMA I/sigma  sd Mn(I)/sd Nmeas  Nref  Ncent FRCBIAS  Nbias
```

--:-- scala_batches_dh_hr_4.log (Text Fill)--L3176--C0--68%

Fig. 10.9 SCALA: data evaluation analysis against resolution for 4-BUDH.

File Edit Options Buffers Tools Help

N	1/resol^2	Dmin	Nmeas	Nref	Ncent	%poss	C%poss	Mlplct	AnoCmpl	AnoFrc	AnoMlt	Rmeas	RmeasO	(Rsym)	PCV	PCVO
$$																
$$																
1	0.039	5.06	32855	9417	1380	95.7	95.7	3.5	93.9	94.1	1.9	0.036	0.037	0.026	0.037	0.041
2	0.078	3.58	59957	16921	1473	97.1	96.6	3.5	93.7	93.7	1.9	0.036	0.034	0.027	0.038	0.039
3	0.117	2.92	81731	22284	1577	99.6	98.0	3.7	98.8	98.8	1.9	0.046	0.043	0.034	0.048	0.049
4	0.156	2.53	97340	26321	1607	99.9	98.7	3.7	99.9	99.9	1.8	0.062	0.060	0.046	0.065	0.068
5	0.195	2.26	110318	29792	1614	100.0	99.1	3.7	99.9	99.9	1.8	0.082	0.080	0.061	0.086	0.092
6	0.234	2.07	120973	32803	1612	100.0	99.3	3.7	100.0	100.0	1.8	0.111	0.109	0.082	0.117	0.127
7	0.273	1.91	131020	35653	1621	100.0	99.4	3.7	100.0	100.0	1.8	0.156	0.157	0.115	0.165	0.185
8	0.312	1.79	137972	38231	1618	100.0	99.5	3.6	99.8	99.8	1.8	0.240	0.243	0.176	0.258	0.288
9	0.352	1.69	144005	40592	1599	100.0	99.6	3.5	99.2	99.2	1.8	0.367	0.368	0.268	0.398	0.439
10	0.391	1.60	150775	42896	1604	99.9	99.7	3.5	98.7	98.7	1.7	0.555	0.557	0.405	0.609	0.666
$$																

```
">&lt;b&gt;For inline graphs use a Java browser&lt;/b&gt;&lt;/applet&gt;

          Overall   1066946  294910  15705   99.7   99.7    3.6   99.0   99.0    1.8   0.075   0.074   0.055   0.078   0.085
                     Nmeas    Nref   Ncent  %poss C%poss Mlplct AnoCmpl AnoFrc AnoMlt  Rmeas  RmeasO  (Rsym)   PCV    PCVO
```

--:-- scala_batches_dh_hr_4.log (Text Fill)--L3255--C0--70%

Fig. 10.10 SCALA: completeness and multiplicity against resolution for 4-BUDH.

to the low R_{merge} value. The log file contains additional information about intensity distributions on the three reciprocal cell axes, which permits one to determine if the relevant axis is a twofold screw axis, or not. Figure 10.11 depicts this information for the a^*-axis. All reflections with an even value for h have large intensity and I/sigI values in average compared to those for odd h values. This means that the extinction rule is fulfilled for a twofold screw axis parallel to the a^*-axis. Similar intensity distributions are found for the b^*- and c^*-axes, giving $P2_12_12_1$ as exact space group. The exact symmetry information must be inserted in the header of the output mtz file from "truncate". This can be done with CCP4 routine "cad".

The data sets collected at the peak and inflection point wavelengths have been evaluated using the same procedure, which gave comparably good reflection statistics. The three data sets must now be scaled together, which was done with

File Edit Options Buffers Tools Help

```
$TABLE: Axial reflections, axis h, Unspecified :
$GRAPHS: I/sigI vs. h:0|    63x0|     22.92:1,4:
     : I vs. h:0|    63x0|  75714.69:1,2:$$

    h         I      sigI    I/sigI            $$ $$

    3       20.       16.     1.266
    4     1367.       75.    18.265
    5      197.       23.     8.593
    6    11467.      550.    20.840
    7      154.       33.     4.750
    8     4257.      210.    20.271
    9      -63.       43.    -1.474
   11      152.       54.     2.798
   12     2568.      137.    18.767
   13      408.       64.     6.405
   14    16373.      789.    20.759
   15      165.       70.     2.345
   16     1814.      118.    15.366
   17      514.       86.     5.989
   18    60272.     2908.    20.725
   19      262.       94.     2.777
   20    27120.     1319.    20.556
   21      376.      103.     3.661
   22     1844.      144.    12.777
   23      219.      118.     1.865
   24    26633.     1279.    20.824
   25     3341.      219.    15.284
   26    21619.     1057.    20.462
   27      -36.      154.    -0.232
   28     2582.      213.    12.134
   29     1729.      194.     8.912
   30    18550.      919.    20.188
   31     -411.      177.    -2.319
   32    37737.     1845.    20.455
   33     -316.      168.    -1.880
   34    68832.     3387.    20.324
   35      747.      166.     4.492
   36    19325.      963.    20.074
   37     2473.      210.    11.782
   38     2845.      215.    13.208
--:--  scala_batches_dh_hr_4.log      (Text Fill)--L3304--C
```

Fig. 10.11 SCALA: axial reflections for axis a^* (h 0 0) for 4-BUDH.

the proper CCP4 routine. The respective input file is shown in Figure 10.12. First, the labels for F, SIGF, DANO and SIGDANO were renamed to the individual values for the respective data sources, suffix RM for remote, suffix INF for inflection, and suffix PK for peak. The actual scaling is done in "scaleit", and the final reflection file is directed to HKLOUT, here into the file "budh_-MAD_scaled.mtz". With this step the data evaluation and reduction has been finished and the phase determination can be started.

The treatment of MIR data sets is similar. The remote data set corresponds to the native one, and the data sets collected at other wavelength of interest conform to the derivative data sets.

```
File  Edit  Options  Buffers  Tools  Help
#!/bin/csh -f
# Scale native and derivative data from MAD after scala
#
setenv REMOTE /home/messersc/budh/dh_hr_4/mosflm/budh_dh_hr_4_trn.mtz
setenv FPRIME /home/messersc/budh/inflection/mosflm/budh_inflection_trn.mtz
setenv FPPRIM /home/messersc/budh/peak/mosflm/budh_peak_trn.mtz
#
# Sort and collect into a single set
#
banner "CAD" >> scal_MAD_mosflm.log
#
cad HKLIN1 ${REMOTE} HKLIN2 ${FPRIME} HKLIN3 ${FPPRIM} \
    HKLOUT tmp1.mtz \
    << EOF-C >> scal_MAD_mosflm.log
TITLE Merged BUDH MAD Data
SYMMETRY 19
RESOLUTION OVERALL 58.0 1.6
SORT H K L
LABIN   FILE 1  E1=F   E2=SIGF   E3=DANO  E4=SIGDANO
CTYP    FILE 1  E1=F   E2=Q      E3=D     E4=Q
LABOUT  File 1  E1=F_RM  E2=SIGF_RM  E3=DANO_RM  E4=SIGDANO_RM

LABIN   FILE 2  E1=F   E2=SIGF   E3=DANO  E4=SIGDANO
CTYP    FILE 2  E1=F   E2=Q      E3=D     E4=Q
LABOUT  File 2  E1=F_INF  E2=SIGF_INF  E3=DANO_INF E4=SIGDANO_INF

LABIN   FILE 3  E1=F   E2=SIGF   E3=DANO  E4=SIGDANO
CTYP    FILE 3  E1=F   E2=Q      E3=D     E4=Q
LABOUT  File 3  E1=F_PK  E2=SIGF_PK  E3=DANO_PK E4=SIGDANO_PK
EOF-C
#
# Scale native and derivative datasets
#
banner "SCALEIT" >>  scal_MAD_mosflm.log
#
setenv HKLIN tmp1.mtz
setenv HKLOUT budh_MAD_scaled.mtz
#
scaleit << EOF-S >> scal_MAD_mosflm.log
TITLE BUDH MAD Data SLS Scaled
RESOLUTION 58.0 1.6
REFINE ANISOTROPIC
GRAPH hkl
CONVERGE NCYC 20

LABIN FP=F_RM   SIGFP=SIGF_RM -
      FPH1=F_RM  SIGFPH1=SIGF_RM  DPH1=DANO_RM   SIGDPH1=SIGDANO_RM -
      FPH2=F_INF  SIGFPH2=SIGF_INF  DPH2=DANO_INF  SIGDPH2=SIGDANO_INF -
      FPH3=F_PK  SIGFPH3=SIGF_PK  DPH3=DANO_PK  SIGDPH3=SIGDANO_PK
END
EOF-S
-u:--  scal_MAD_mosflm1.com      (Lisp Interaction)--L1--C0--Top----------
```

Fig. 10.12 Input file for scaling of remote, inflection and peak datasets for 4-BUDH.

10.3
A Complex Example of Space Group Determination

For 4-BUDH the space group could be determined unambiguously from the autoindexing and scaling procedure and detected extinctions. We will now discuss a rather complicated case of a space group determination. This was met when solving the crystal structure of the catalytic domain of human atypical protein kinase C-iota (PKC-iota) (Messerschmidt et al., 2005). We will also use this example to illustrate the application of the method of Molecular Replacement.

The protein had been produced from SF9 insect cells infected with recombinant baculovirus. The protein construct had an N-terminal His_6-tag with a TEV cleavage

File Edit Search Preferences Shell Macro Windows Help

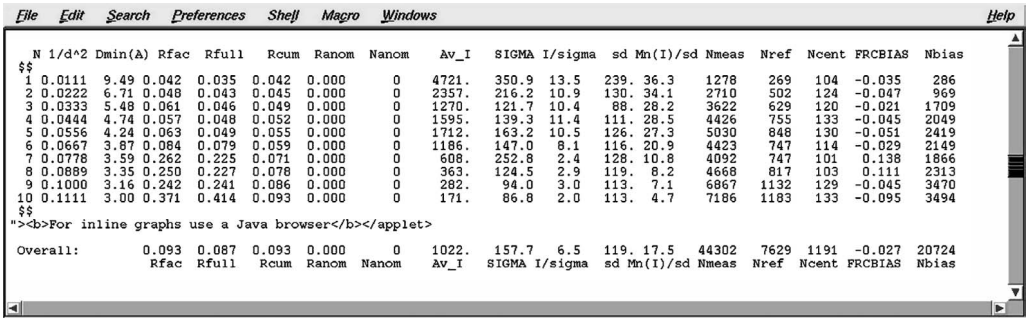
```
List of possible Laue groups, sorted on penalty index.
The lower the PENALTY, the better
No PENALTY LATT      a        b        c     alpha  beta gamma  Possible spacegroups

 44 999    oI   137.85    77.96   136.92    74.1  58.1  74.9   I23,I213,I432,I4132
 43 995    hR   175.18   137.78    77.68    75.1  90.1  80.5   H3,H32 (hexagonal settings of R3 and R32)
 42 873    oF   137.12   174.72   139.67   122.0  68.6 122.4   F23,F432,F4132
 41 850    tI   136.92   139.67    76.92    73.9 106.2 121.4   I4,I41,I422,I4122
 40 740    tI   137.85    77.96   136.92    74.1  58.1  74.9   I4,I41,I422,I4122
 39 732    hP    77.68   114.35    76.92    90.0 119.4  91.2   P3,P31,P32,P312,P321,P3112,P3121,P3212,P3221
                                                               P6,P61,P65,P62,P64,P63,P622,P6122,P6522,P6222,P6422,P6322
 38 680    oI    76.92    77.68   240.16    98.4  99.3 119.4   I222,I212121
 37 680    tI    76.92    77.68   240.16    81.6  80.7 119.4   I4,I41,I422,I4122
 36 610    tI   137.12   137.78    77.68    75.1  74.1  58.3   I4,I41,I422,I4122
 35 608    oI    76.92   136.92   139.67    58.6  73.9  73.8   I222,I212121
 34 602    oI    77.68   137.78   137.12    58.3  74.1  75.1   I222,I212121
 33 543    hR    76.92    77.96   366.60    89.2 108.3 119.8   H3,H32 (hexagonal settings of R3 and R32)
 32 521    oF    76.92   135.30   241.25    88.7 108.6  90.2   F222
 31 481    cP    76.92    77.68   114.35    91.2  90.0 119.4   P23,P213,P432,P4232,P4332,P4132
 30 476    tP    77.68   114.35    76.92    90.0 119.4  91.2   P4,P41,P42,P43,P422,P4212,P4122,P41212,P4222,P42212,P4322,P43212
 29 473    hR   133.50   137.78   137.12   121.7  90.2 118.1   H3,H32 (hexagonal settings of R3 and R32)
 28 472    hR    77.96   137.78   176.84   100.4  88.8 107.1   H3,H32 (hexagonal settings of R3 and R32)
 27 467    oC   136.92   139.55    76.92    74.1 106.2  68.4   C222,C2221
 26 467    mC   136.92   139.55    76.92    74.1 106.2  68.4   C2
 25 466    mC   139.55   136.92    76.92   106.2 105.9 111.6   C2
 24 409    oC    77.68   240.03    76.92    80.9 119.4 107.7   C222,C2221
 23 408    mC   240.03    77.68    76.92   119.4  99.1  72.3   C2
 22 403    oC    77.68   240.03    76.92    80.9 119.4 107.7   C222,C2221
 21 403    mC   240.03    77.68    76.92   119.4  99.1  72.3   C2
 20 401    mC    77.68   240.03    77.96    98.4 120.8  72.3   C2
 19 397    mC    76.92   241.25    77.68    97.9 119.4  71.4   C2
 18 396    mI    77.68   240.03    76.92    99.1 119.4  72.3   C2 (transformed from I2)
 17 282    oF    77.96   133.50   240.16    89.5 107.8  89.4   F222
 16 278    mC   135.30    76.92   136.92    73.8 118.3  89.8   C2
 15 273    mC   133.50    77.96   137.85    74.9 119.4  89.4   C2
 14 152    tP    76.92    77.68   114.35    91.2  90.0 119.4   P4,P41,P42,P43,P422,P4212,P4122,P41212,P4222,P42212,P4322,P43212
 13 146    oP    76.92    77.68   114.35    91.2  90.0 119.4   P222,P2221,P21212,P212121
 12 143    mP    77.68    76.92   114.35    90.0  91.2 119.4   P2,P21
 11 138    mP    77.68    76.92   114.35    90.0  91.2 119.4   P2,P21
 10  19    mC    76.92   135.30   114.35    88.7  90.0  89.8   C2
  9  16    hP    76.92    77.68   114.35    91.2  90.0 119.4   P3,P31,P32,P312,P321,P3112,P3121,P3212,P3221
                                                               P6,P61,P65,P62,P64,P63,P622,P6122,P6522,P6222,P6422,P6322
  8  14    mC    77.96   133.50   114.35    90.7  91.1  89.4   C2
  7  14    oC    77.96   133.50   114.35    90.7  91.1  89.4   C222,C2221
  6  11    oC    76.92   135.30   114.35    88.7  90.0  90.2   C222,C2221
  5  10    mC   135.30    76.92   114.35    90.0  91.3  89.8   C2
  4   9    mP    76.92   114.35    77.68    91.2 119.4  90.0   P2,P21
  3   2    mC   135.30    76.92   114.35    90.0  91.3  89.8   C2
  2   0    aP    76.92    77.68   114.35    88.8  90.0  60.6   P1
  1   0    aP    76.92    77.68   114.35    91.2  90.0 119.4   P1
No PENALTY SDCELL FRACN LATT      a        b        c     alpha  beta gamma  Possible spacegroups

Suggested Solution:  9  P3
          penalty: 16
             cell:  76.917   77.680  114.353  91.17  89.97 119.44
  regularized cell:  77.299   77.299  114.353  90.00  90.00 120.00
          Symmetry: hP (Primitive Hexagonal)
Select a solution AND a spacegroup from list above (eg 3 p42) or 0 to abandon:
```

Fig. 10.13 MOSFLM: Output of autoindexing for PKC-iota.

site, which was cleaved off before crystallization. Crystals could be obtained as sitting drops by vapor diffusion with PEG 400 (24–34%) as the main precipitant. An intensity data set was collected to a resolution of 2.8 Å at the synchrotron beamline PX of the Swiss Light Source at the Paul Scherrer Institute, Villigen, Switzerland using a MAR CCD 165 detector (MarResearch, Norderstedt, Germany). The data were also evaluated with program MOSFLM, and the output of the autoindexing is shown in Figure 10.13. The suggested solution is number 9, which has a reasonably low penalty and the highest symmetry among this group. The problem is that this solution comprises a large number of possible space groups. They belong to four different Laue groups ($\bar{3}, \bar{3}m, 6/m, 6/mmm$). Generally, one has to evaluate the data set for the four Laue groups, each with the space group of the lowest symmetry of the respective Laue group. Here, we would run MOSFLM and SCALA for space groups *P*3, *P*3(1)2, *P*6 and *P*622. We have also to include *P*32(1) because the twofold axes perpendicular to the threefold axis may have two different orientations 30° apart from each other. The best R_{merge} was obtained for *P*32(1), as shown in Figure 10.14. The value of 0.093 is satisfying and is half of the values which were received for the other four possibilities. The corresponding listing for the



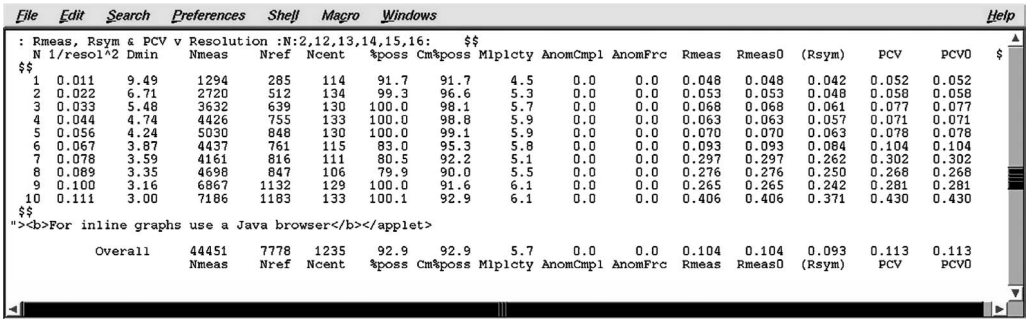
File Edit Search Preferences Shell Macro Windows Help

N	1/d^2	Dmin(A)	Rfac	Rfull	Rcum	Ranom	Nanom	Av_I	SIGMA I/sigma		sd Mn(I)/sd		Nmeas	Nref	Ncent	FRCBIAS	Nbias
$$																	
1	0.0111	9.49	0.042	0.035	0.042	0.000	0	4721.	350.9	13.5	239.	36.3	1278	269	104	-0.035	286
2	0.0222	6.71	0.048	0.043	0.045	0.000	0	2357.	216.2	10.9	130.	34.1	2710	502	124	-0.047	969
3	0.0333	5.48	0.061	0.046	0.049	0.000	0	1270.	121.7	10.4	88.	28.2	3622	629	120	-0.021	1709
4	0.0444	4.74	0.057	0.048	0.052	0.000	0	1595.	139.3	11.4	111.	28.5	4426	755	133	-0.045	2049
5	0.0556	4.24	0.063	0.049	0.055	0.000	0	1712.	163.2	10.5	126.	27.3	5030	848	130	-0.051	2419
6	0.0667	3.87	0.084	0.079	0.059	0.000	0	1186.	147.0	8.1	116.	20.9	4423	747	114	-0.029	2149
7	0.0778	3.59	0.262	0.225	0.071	0.000	0	608.	252.8	2.4	128.	10.8	4092	747	101	0.138	1866
8	0.0889	3.35	0.250	0.227	0.078	0.000	0	363.	124.5	2.9	119.	8.2	4668	817	103	0.111	2313
9	0.1000	3.16	0.242	0.241	0.086	0.000	0	282.	94.0	3.0	113.	7.1	6867	1132	129	-0.045	3470
10	0.1111	3.00	0.371	0.414	0.093	0.000	0	171.	86.8	2.0	113.	4.7	7186	1183	133	-0.095	3494
$$																	

">For inline graphs use a Java browser</applet>

Overall:	0.093	0.087	0.093	0.000	0	1022.	157.7	6.5	119.	17.5	44302	7629	1191	-0.027	20724
	Rfac	Rfull	Rcum	Ranom	Nanom	Av_I	SIGMA I/sigma		sd Mn(I)/sd		Nmeas	Nref	Ncent	FRCBIAS	Nbias

Fig. 10.14 R_{merge} against resolution for the data set of PKC-iota for symmetry *P*32(1).



File Edit Search Preferences Shell Macro Windows Help

: Rmeas, Rsym & PCV v Resolution :N:2,12,13,14,15,16: $$

N	1/resol^2	Dmin	Nmeas	Nref	Ncent	%poss	Cm%poss	Mlplcty	AnomCmpl	AnomFrc	Rmeas	RmeasO	(Rsym)	PCV	PCVO	$
$$																
1	0.011	9.49	1294	285	114	91.7	91.7	4.5	0.0	0.0	0.048	0.048	0.042	0.052	0.052	
2	0.022	6.71	2720	512	134	99.3	96.6	5.3	0.0	0.0	0.053	0.053	0.048	0.058	0.058	
3	0.033	5.48	3632	639	130	100.0	98.1	5.7	0.0	0.0	0.068	0.068	0.061	0.077	0.077	
4	0.044	4.74	4426	755	133	100.0	98.8	5.9	0.0	0.0	0.063	0.063	0.057	0.071	0.071	
5	0.056	4.24	5030	848	130	100.0	99.1	5.9	0.0	0.0	0.070	0.070	0.063	0.078	0.078	
6	0.067	3.87	4437	761	115	83.0	95.3	5.8	0.0	0.0	0.093	0.093	0.084	0.104	0.104	
7	0.078	3.59	4161	816	111	80.5	92.2	5.1	0.0	0.0	0.297	0.297	0.262	0.302	0.302	
8	0.089	3.35	4698	847	106	79.9	90.0	5.5	0.0	0.0	0.276	0.276	0.250	0.268	0.268	
9	0.100	3.16	6867	1132	129	100.0	91.6	6.1	0.0	0.0	0.265	0.265	0.242	0.281	0.281	
10	0.111	3.00	7186	1183	133	100.1	92.9	6.1	0.0	0.0	0.406	0.406	0.371	0.430	0.430	
$$																

">For inline graphs use a Java browser</applet>

Overall	44451	7778	1235	92.9	92.9	5.7	0.0	0.0	0.104	0.104	0.093	0.113	0.113	
	Nmeas	Nref	Ncent	%poss	Cm%poss	Mlplcty	AnomCmpl	AnomFrc	Rmeas	RmeasO	(Rsym)	PCV	PCVO	

Fig. 10.15 Completeness against resolution for the data set of PKC-iota for symmetry *P*32(1).

completeness of the data set is depicted in Figure 10.15. This is sufficient, with an overall of 92.9%. Some reflections had to be omitted because of ice rings documented by lower values of completeness between 3.87 and 3.35 Å resolution.

Finally, one must check for extinctions of the *c** axis. The relevant intensity list is displayed in Figure 10.16. It can be clearly seen that only reflections with $l = 3n$ are present. Reflection (0, 0, 31) has obviously been determined incorrectly, probably due to the ice rings on the diffraction images. The observed extinctions indicate a threefold screw axis parallel to the *c** axis. From the extinctions, we cannot distinguish between a 3_1 or 3_2 axis. Therefore, we have space groups $P3_12(1)$ or $P3_22(1)$ as possibilities at the end of this analysis. The final decision can be made during the phase determination, where the wrong space group delivers incorrect phases. With this example, the problem can be used during the molecular replacement step. The molecular replacement is calculated for both space groups and the correct solution is gained for the proper space group.

Although the data set showed diffraction to 2.8 Å resolution, the data have been evaluated to 3.0 Å resolution only, taking into account a threshold of 2 for I/sigma(I).

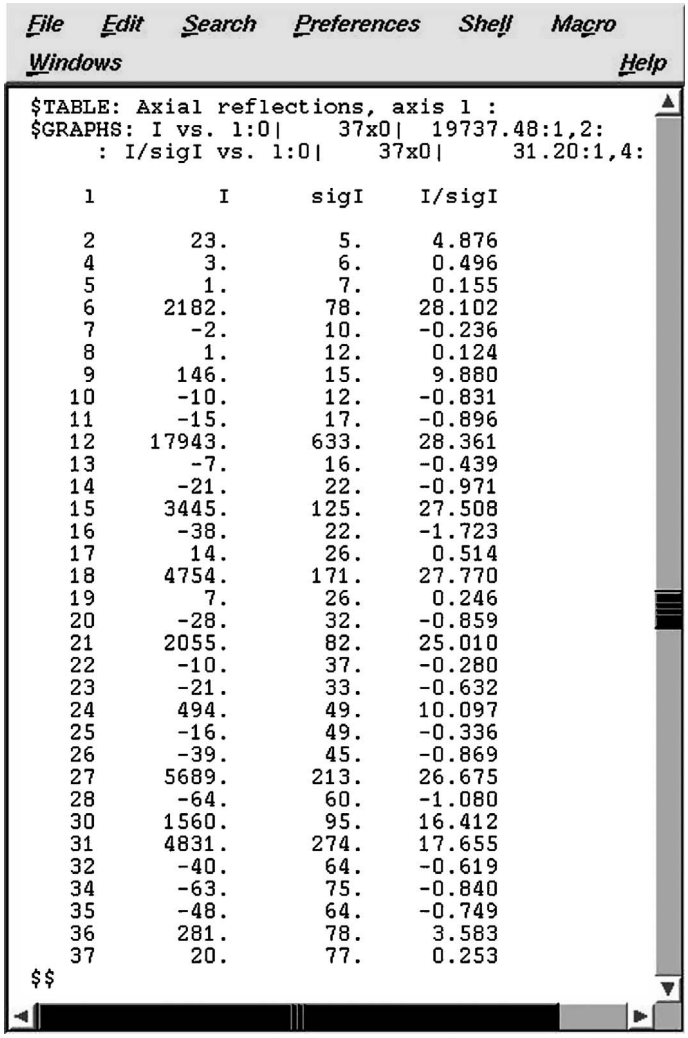


Fig. 10.16 SCALA: axial reflections for axis c^* (0 0 *l*) for PKC-iota.

References

Martins, B. M., Dobbek, H., Çinkaya, I., Buckel, W., Messerschmidt, A., *Proc. Natl. Acad. Sci. USA* **2004**, *44*, 15645–15649.

Messerschmidt, A., Macieira, S., Velarde, M., Bädeker, M., Benda, C., Jestel, A., Brandstetter, H., Neuefeind, T., Blaesse, M., *J. Mol. Biol.* **2005**, *352*, 918–931.

11
Determination of Anomalous Scatterer or Heavy Atom Positions

The methods used to determine the positions of the anomalous scatterers or heavy atoms are the same, and their theoretical basis has been explained in Sections 5.3.1 and 5.3.2. Here, we use the structure analysis of 4-BUDH (Martins et al., 2004) to illustrate the identification of its 16 Fe-sites (four [4Fe-4S] clusters per asymmetric unit), whose anomalous scattering effect at the Fe K_α-absorption edge has been used. Both direct and vector verification methods will be discussed and compared.

11.1
Application of Direct Methods

Shake and Bake (SnB) (Weeks and Miller, 1999) has been used as the direct methods program. When one starts SnB, a GUI is opened with information

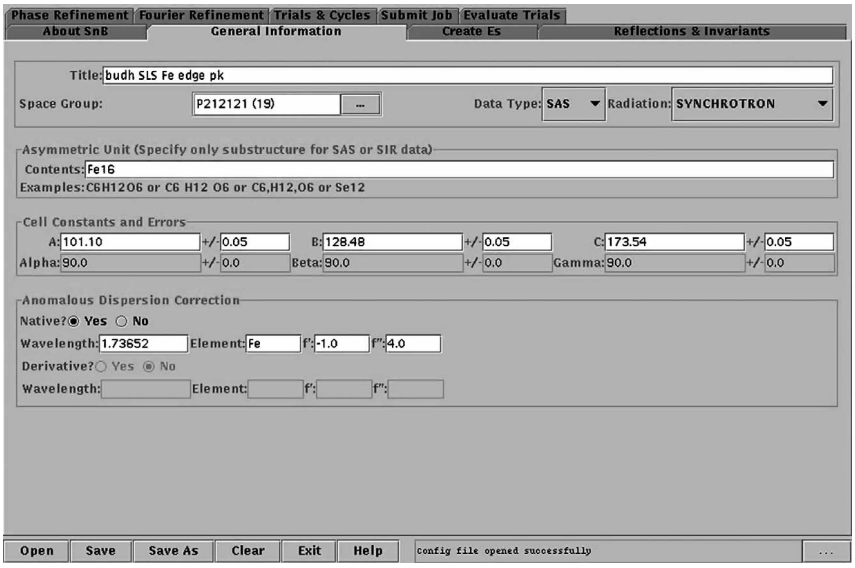


Fig. 11.1 SnB Input GUI: General Information.

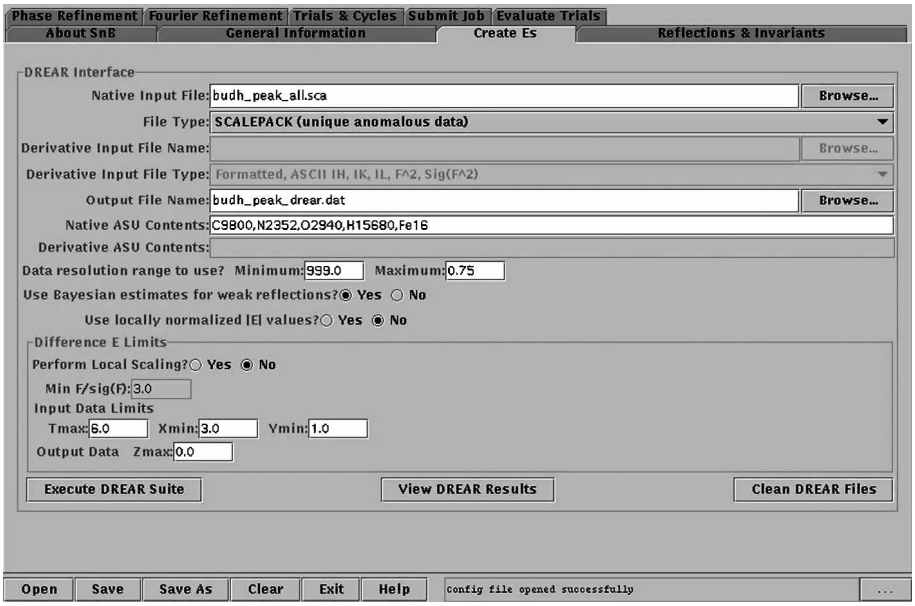


Fig. 11.2 SnB Input GUI: Create Es.

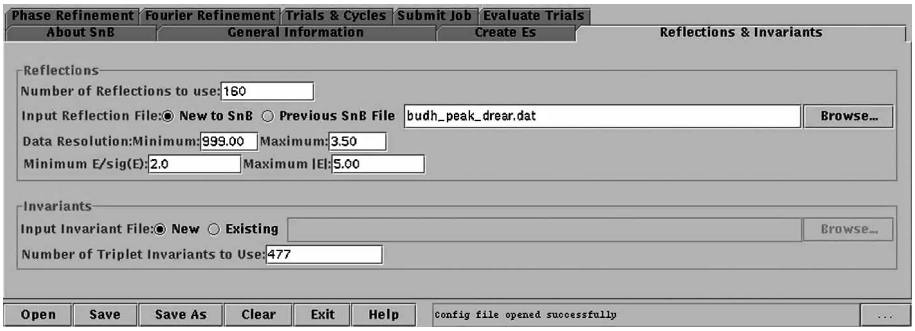


Fig. 11.3 SnB Input GUI: Reflections & Invariants.

about SnB. First, we choose the screen tab "General information", as displayed in Figure 11.1. A meaningful title, the space group, cell constants with errors and the number of anomalous scatterers per asymmetric unit (Fe 16 in this case) must be entered. Three choices are possible for Data Type: Basic, SAS and SIR. We select SAS, which stands for single anomalous scattering, because we use the anomalous differences collected at the Fe K_α-absorption edge with maximum f′. "Synchrotron" must be indicated for Radiation. Finally, one must supply the element of the anomalous scatterer, the f′ and f″ values and the wavelength at which the data set has been registered. Values of f′ and f″ for the re-

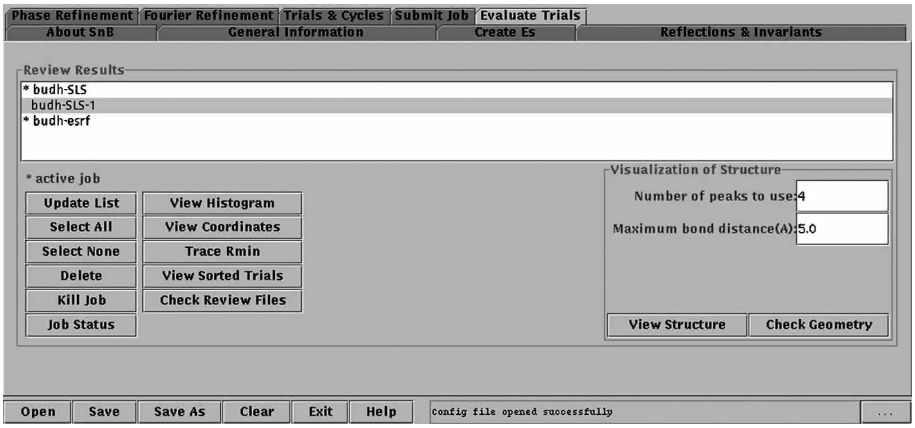


Fig. 11.4 SnB Output GUI: Evaluate Trials.

spective wavelengths can be obtained from the X-ray fluorescence scan, which is usually done at the synchrotron beamline before the MAD data collection.

Before one can run SnB the normalized structure factors *E* of the anomalous differences must be calculated. This is done by program DREAR (Blessing and Smith, 1999) and the input GUI is supplied by choosing screen tab "Create Es" (Fig. 11.2). SCALEPACK (unique anomalous data) has been selected as File Type and the relevant reflection file is "budh_peak_all.sca", the output file from Scalepack for the respective wavelength. If the diffraction data has been stored

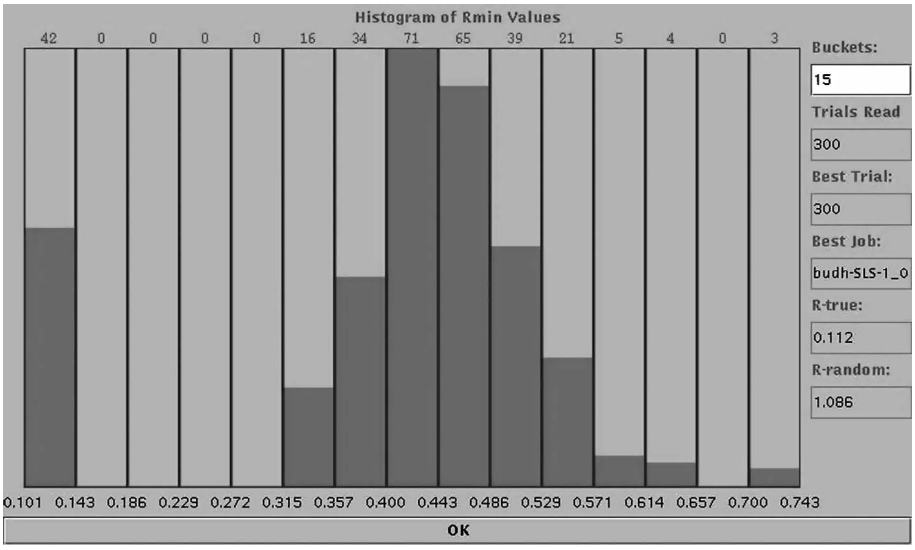
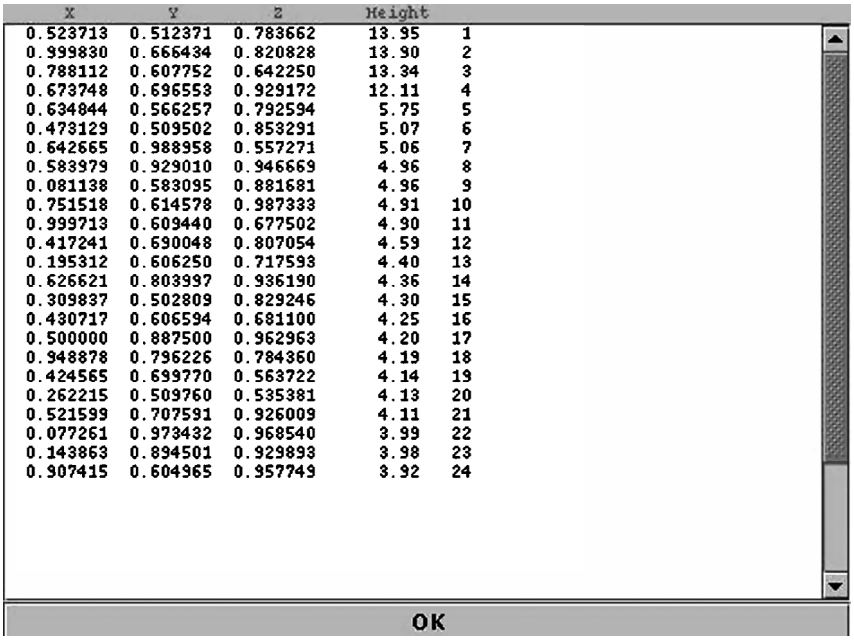


Fig. 11.5 SnB Output GUI: Histogram of minimal function.



X	Y	Z	Height	
0.523713	0.512371	0.783662	13.95	1
0.999830	0.666434	0.820828	13.90	2
0.788112	0.607752	0.642250	13.34	3
0.673748	0.696553	0.929172	12.11	4
0.634844	0.566257	0.792594	5.75	5
0.473129	0.509502	0.853291	5.07	6
0.642665	0.988958	0.557271	5.06	7
0.583979	0.929010	0.946669	4.96	8
0.081138	0.583095	0.881681	4.96	9
0.751518	0.614578	0.987333	4.91	10
0.999713	0.609440	0.677502	4.90	11
0.417241	0.690048	0.807054	4.59	12
0.195312	0.606250	0.717593	4.40	13
0.626621	0.803997	0.936190	4.36	14
0.309837	0.502809	0.829246	4.30	15
0.430717	0.606594	0.681100	4.25	16
0.500000	0.887500	0.962963	4.20	17
0.948878	0.796226	0.784360	4.19	18
0.424565	0.699770	0.563722	4.14	19
0.262215	0.509760	0.535381	4.13	20
0.521599	0.707591	0.926009	4.11	21
0.077261	0.973432	0.968540	3.99	22
0.143863	0.894501	0.929893	3.98	23
0.907415	0.604965	0.957749	3.92	24

OK

Fig. 11.6 SnB Output GUI: Coordinates.

in a CCP4 MTZ-file, one may create a Scalepack formatted file with the OUTPUT SCAL keyword in the CCP4 routine MTZ2VARIOUS. The native contents of the ASU (asymmetric unit) has been approximated by the relations C=5R, N=1.2R, O=1.5R and H=8R, where R equals the number of protein residues in the ASU (here, 4×490=1960). The default values have been taken for the other parameters.

If the DREAR job has been finished, one can move to the screen tab "Reflections & Invariants" (Fig. 11.3). For the determination of the positions of the anomalous scatterers lower to medium resolution reflections should be included only. Here, a maximum limit of 3.5 Å has been chosen. The adjustment of the other parameters needs some test runs. The number of reflections to be used and the E/sig(E) are correlated. In this example, a value of 2.0 worked well and corresponded to 160 reflections. The possible triple invariants are calculated and their number is compared with the input value. If this number is less than the input value, the program stops and an error message is written to a file in the working directory. This file contains the number of triple invariants calculated for the given parameters and must be entered in the relevant parameter field of the "Reflections & Invariants" screen tab.

Default values have been taken from the other screen tabs except for "Trials & Cycles" where a number of 300 trials has been selected. It is the normal case to run the SnB job interactively, but in the screen tab "Submit job" one can also write data files suitable to run SnB in batch mode. The screen tab "Evaluate

Fig. 11.7 SnB Output GUI: Visualization.

Trials" provides several means to check the results of the SnB job both during the job process or after completion of the job. In Figure 11.4 the job "budh-SLS-1" has been chosen. As the resolution is 3.5 Å only, the individual Fe-atoms in the Fe–S clusters will not be resolved. Therefore, the expected number of Fe sites will be four for one homotetramer, and the number of peaks to use has been set to four in the input field for "Visualization of Structure". We will not discuss all options of this screen tab. The "View Histogram" is very useful. The final histogram is displayed in Figure 11.5.

The histogram in Figure 11.5 contains the distribution of the minimal function for all trials. A pronounced bimodal distribution is a reliable indication for a correct solution, which are in the distribution with the lowest minimal function values. The list of coordinates (Fig. 11.6) for the best solution can be produced by pressing the "View Coordinates" button in the "Evaluate Trials" button. As expected, there are four sites with almost equal heights and clearly separated from the heights of the following sites. Pressing the button "View Structure" activates a window showing a ball-and-stick model based on the peak file and the maximum distance specified (Fig. 11.7). One can edit the model, identify atoms, and save the revised model in a file. Here, the four Fe-sites are displayed and the determination of the inter-site distances with the "Distance" button, and clicking the sites of interest may be very useful. The coordinates of the four Fe-sites (see top four peaks in list of Fig. 11.6) have y- and z-coordinates all be-

tween 0.5 and 1.0. For the later use, the Fe-sites have been referred to a unit cell origin at (0, 0.5, 0.5), what means that their y- and z-coordinates have been subtracted by 0.5, respectively. Origin shifts of 0.5 do not affect the space group symmetry in space group $P2_12_12_1$, the space group of the 4-BUDH example.

11.2 Vector Verification Methods

We explain the application of the vector verification methods for our 4-BUDH example by means of the Real Space Patterson Search (RSPS) program of Knight (2000), which is part of the CCP4 suite (CCP4, 1994). The principles of vector verification have been described in Section 5.3.1. We illustrate the running of the CCP4 programs by using the CCP4i GUI. If you type "ccp4i" on your computer where the CCP4 system has been installed, the GUI for the control center of CCP4i is opened (Fig. 11.8). We choose the button "FFT for Patterson" from the program list because the anomalous difference Patterson map must be calculated before we can start the RSPS routine. CCP4i opens a window for entering the input parameters for the Patterson map calculation (Fig. 11.9). The input MTZ-file is "budh_MAD_scaled.mtz" and the label DANO_PK, which stands for the anomalous difference of the f′ wavelength, must be selected. The map has been calculated to a maximum resolution of 3.5 Å.

RSPS is now started by pressing the program list button "RSPS". In the respective input window (Fig. 11.10) we select the options "Get list of potential heavy atoms from scan of Patterson map" and "To analyze sites find sets of sites with good cross vectors", and assign files for the input Patterson map and the output coordinates of sites. The sigma values for picking peaks and sites have been set to 1.5. The output log-file can be invoked from the CCP4i control center. The part for the peak search of the anomalous difference Patterson map is shown in Figure 11.11. An inspection of this peak list allows some conclusions

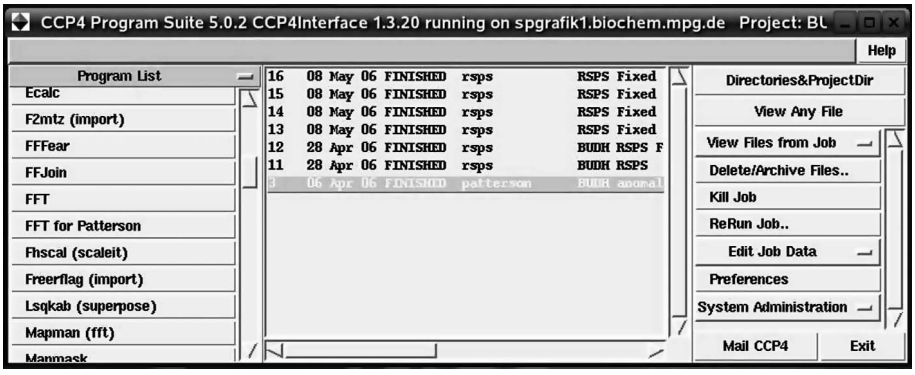


Fig. 11.8 CCP4i control center.

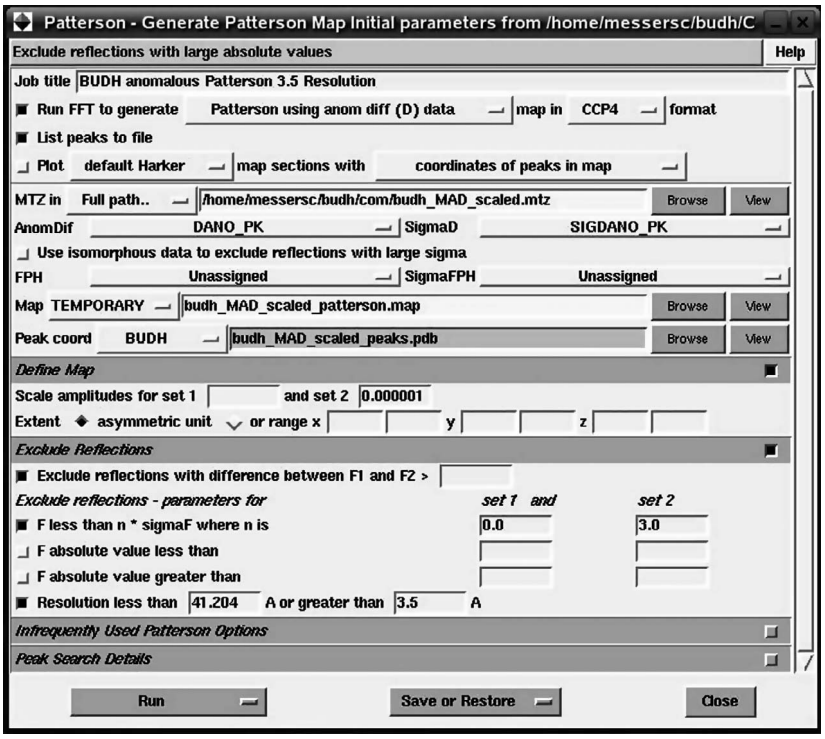


Fig. 11.9 CCP4i: generate Patterson parameters input for 4-BUDH.

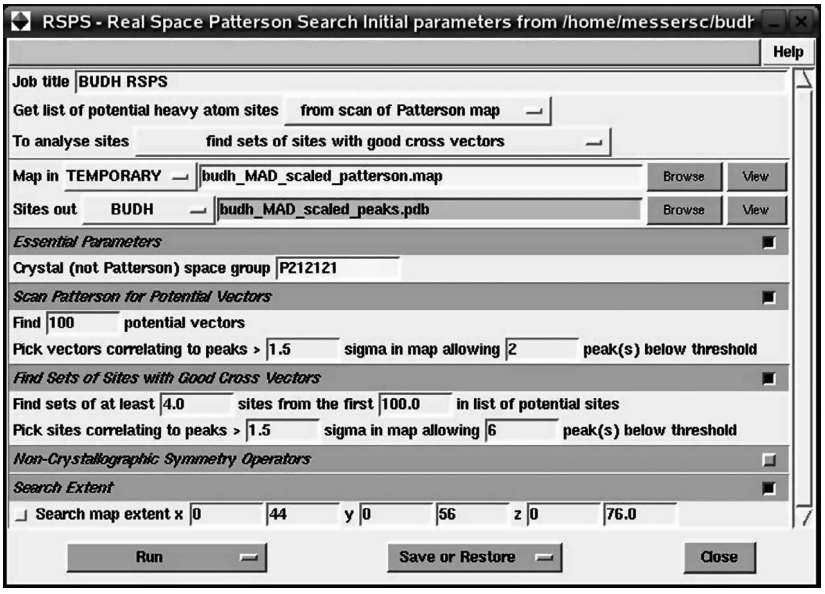


Fig. 11.10 CCP4i: RSPS parameters input for finding sets of sites with good cross vectors for 4-BUDH (first run).

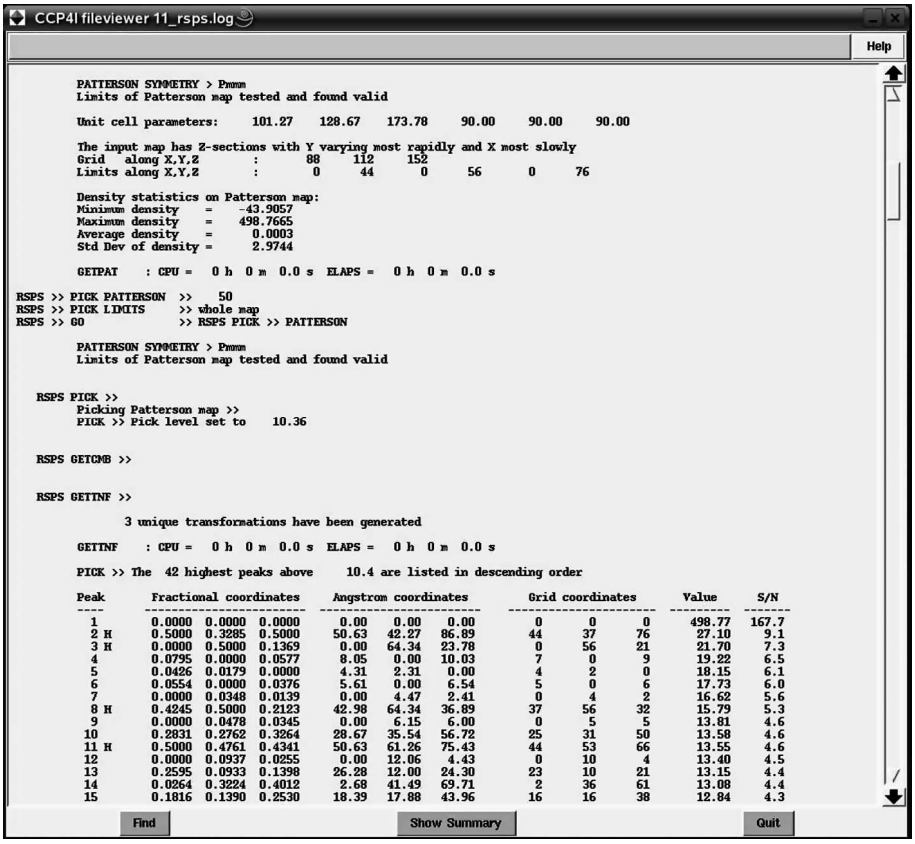
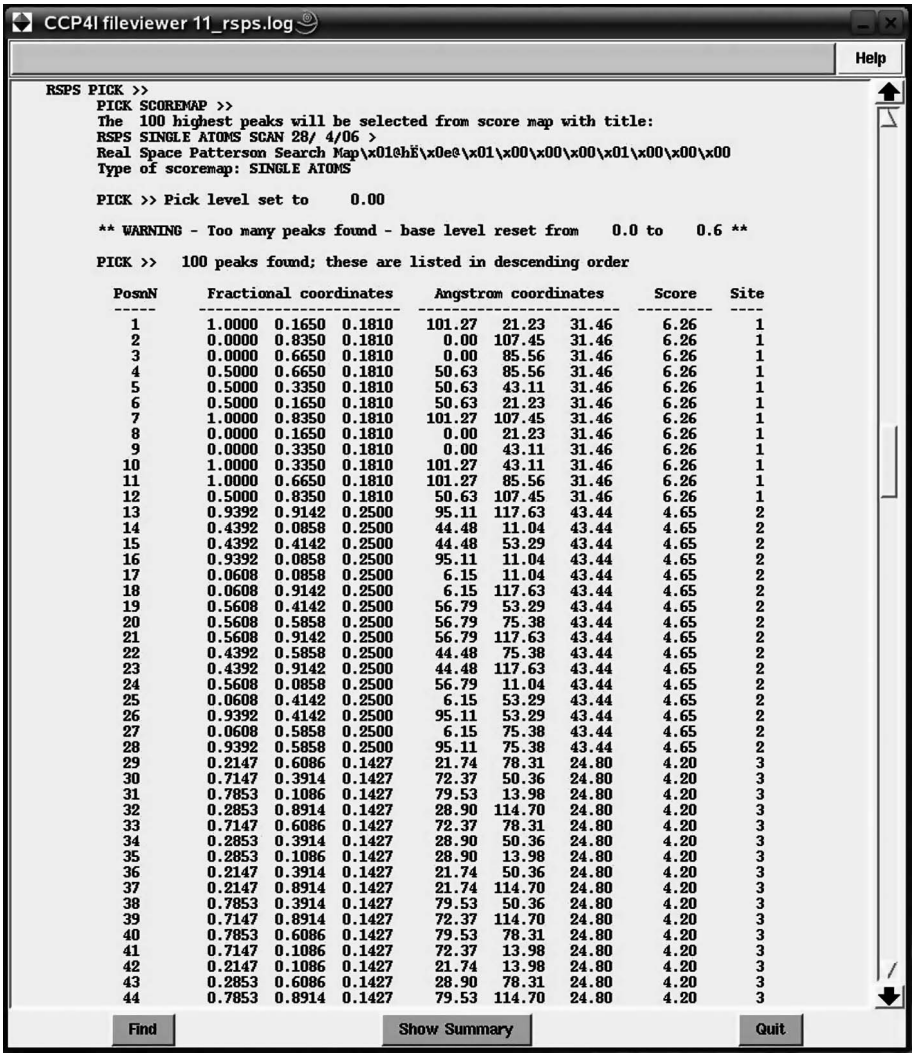


Fig. 11.11 Output for PICK PATTERSON of RSPS for 4-BUDH (first run).

to be drawn about the quality of the Patterson map. At this low resolution one expects heavy atom vectors with a height of 1/16 of the origin peak, and they should not be close to the origin. This is quite well satisfied for the top peaks of the list. Four of these are Harker vectors and have been assigned by the label "H". Next, a SINGLE ATOM SCAN is run according to Eq. (5.48). The results are displayed in Figure 11.12. The three top sites are shown, and they consist of a set of 12 positions, which correspond to the 12 different possible origins in this symmetry group. This CCP4i run additionally invokes the GETSETS routine, which searches for sets of positions. This is performed by looking at the cross-vectors between all pairs of atoms (and their symmetry mates) in the list. As cross-vectors are used, sets of consistent sites are related to the same origin and constitute a solution of our problem.

The results of the GETSETS routine for 4-BUDH are displayed in Figure 11.13. We obtain the expected set of four sites. The score table giving the score for the Harker and cross vectors generated by these positions shows significant



CCP4I fileviewer 11_rsps.log

Help

```
RSPS PICK >>
      PICK SCOREMAP >>
      The  100 highest peaks will be selected from score map with title:
      RSPS SINGLE ATOMS SCAN 28/ 4/06 >
      Real Space Patterson Search Map\x01@hË\x0e@\x01\x00\x00\x00\x01\x00\x00\x00
      Type of scoremap: SINGLE ATOMS

      PICK >> Pick level set to     0.00

      ** WARNING - Too many peaks found - base level reset from    0.0 to    0.6 **

      PICK >>   100 peaks found; these are listed in descending order
```

PosnN	Fractional coordinates			Angstrom coordinates			Score	Site
1	1.0000	0.1650	0.1810	101.27	21.23	31.46	6.26	1
2	0.0000	0.8350	0.1810	0.00	107.45	31.46	6.26	1
3	0.0000	0.6650	0.1810	0.00	85.56	31.46	6.26	1
4	0.5000	0.6650	0.1810	50.63	85.56	31.46	6.26	1
5	0.5000	0.3350	0.1810	50.63	43.11	31.46	6.26	1
6	0.5000	0.1650	0.1810	50.63	21.23	31.46	6.26	1
7	1.0000	0.8350	0.1810	101.27	107.45	31.46	6.26	1
8	0.0000	0.1650	0.1810	0.00	21.23	31.46	6.26	1
9	0.0000	0.3350	0.1810	0.00	43.11	31.46	6.26	1
10	1.0000	0.3350	0.1810	101.27	43.11	31.46	6.26	1
11	1.0000	0.6650	0.1810	101.27	85.56	31.46	6.26	1
12	0.5000	0.8350	0.1810	50.63	107.45	31.46	6.26	1
13	0.9392	0.9142	0.2500	95.11	117.63	43.44	4.65	2
14	0.4392	0.0858	0.2500	44.48	11.04	43.44	4.65	2
15	0.4392	0.4142	0.2500	44.48	53.29	43.44	4.65	2
16	0.9392	0.0858	0.2500	95.11	11.04	43.44	4.65	2
17	0.0608	0.0858	0.2500	6.15	11.04	43.44	4.65	2
18	0.0608	0.9142	0.2500	6.15	117.63	43.44	4.65	2
19	0.5608	0.4142	0.2500	56.79	53.29	43.44	4.65	2
20	0.5608	0.5858	0.2500	56.79	75.38	43.44	4.65	2
21	0.5608	0.9142	0.2500	56.79	117.63	43.44	4.65	2
22	0.4392	0.5858	0.2500	44.48	75.38	43.44	4.65	2
23	0.4392	0.9142	0.2500	44.48	117.63	43.44	4.65	2
24	0.5608	0.0858	0.2500	56.79	11.04	43.44	4.65	2
25	0.0608	0.4142	0.2500	6.15	53.29	43.44	4.65	2
26	0.9392	0.4142	0.2500	95.11	53.29	43.44	4.65	2
27	0.0608	0.5858	0.2500	6.15	75.38	43.44	4.65	2
28	0.9392	0.5858	0.2500	95.11	75.38	43.44	4.65	2
29	0.2147	0.6086	0.1427	21.74	78.31	24.80	4.20	3
30	0.7147	0.3914	0.1427	72.37	50.36	24.80	4.20	3
31	0.7853	0.1086	0.1427	79.53	13.98	24.80	4.20	3
32	0.2853	0.8914	0.1427	28.90	114.70	24.80	4.20	3
33	0.7147	0.6086	0.1427	72.37	78.31	24.80	4.20	3
34	0.2853	0.3914	0.1427	28.90	50.36	24.80	4.20	3
35	0.2853	0.1086	0.1427	28.90	13.98	24.80	4.20	3
36	0.2147	0.3914	0.1427	21.74	50.36	24.80	4.20	3
37	0.2147	0.8914	0.1427	21.74	114.70	24.80	4.20	3
38	0.7853	0.3914	0.1427	79.53	50.36	24.80	4.20	3
39	0.7147	0.8914	0.1427	72.37	114.70	24.80	4.20	3
40	0.7853	0.6086	0.1427	79.53	78.31	24.80	4.20	3
41	0.7147	0.1086	0.1427	72.37	13.98	24.80	4.20	3
42	0.2147	0.1086	0.1427	21.74	13.98	24.80	4.20	3
43	0.2853	0.6086	0.1427	28.90	78.31	24.80	4.20	3
44	0.7853	0.8914	0.1427	79.53	114.70	24.80	4.20	3

Find | Show Summary | Quit

Fig. 11.12 Top of the peak list from SINGLE ATOM SCAN of RSPS for 4-BUDH (first run).

scores for all vectors, which are of the same magnitude. It may happen that the GETSETS routine does not deliver satisfying results. In this case, a cross-vector scan with one or more known sites fixed according to Eq. (5.49) may be performed (MORE ATOMS SCAN in RSPS). The relevant input window for 4-BUDH is given in Figure 11.14. The potential heavy-atom sites are read from the output file "budh_MAD_scaled_peaks.pdb" of the first run of RSPS. The found peaks are written to file "budh_MAD_peaks_out.pdb". One site with coordinates (0.9995, 0.1649, 0.1809) has been fixed. The top of the peak list of the

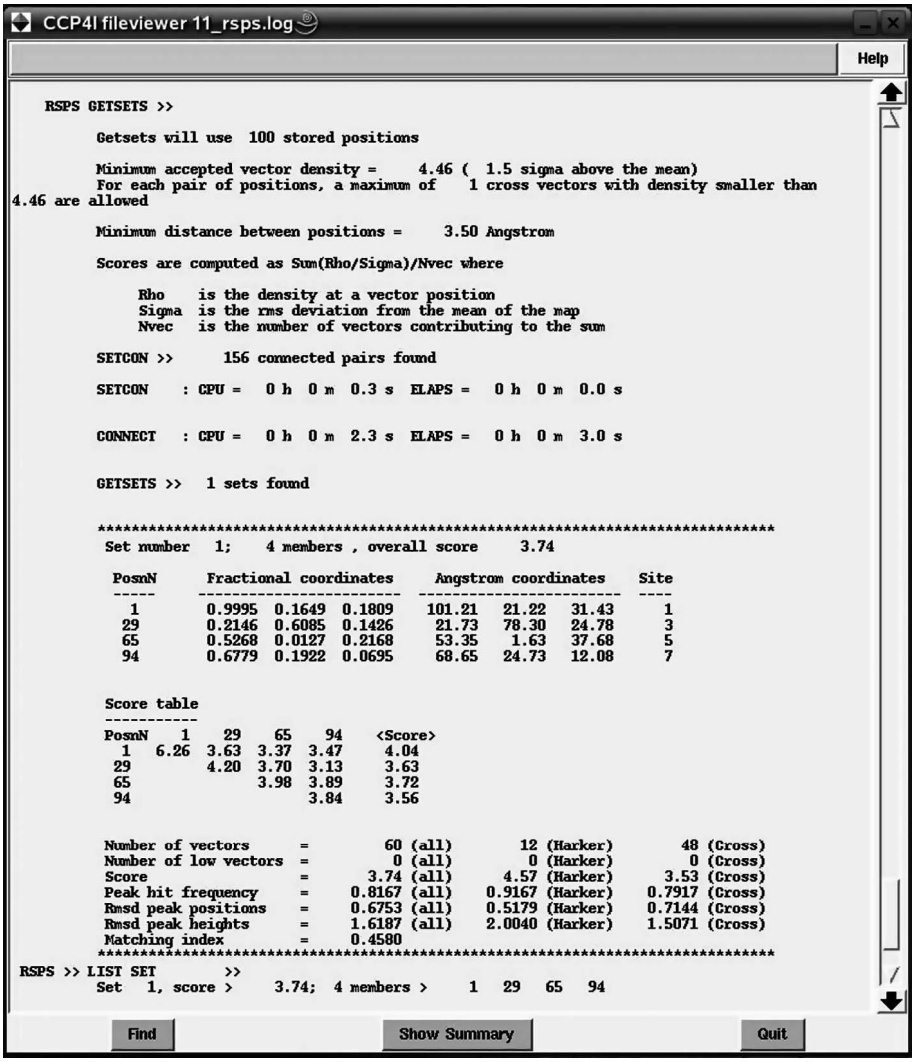


Fig. 11.13 Output of GETSETS of RSPS for 4-BUDH (first run).

score map is displayed in Figure 11.15. We are looking for three further sites, and these are on top of the list. However, they are ambiguous with two possibilities for each site.

A comparison of the four sites from GETSETS with the list shows that they are contained in this list. The set from GETSETS is related to a common origin, whereas the results from the second run are still ambiguous. This can be overcome by using two fixed sites, but this is not shown here.

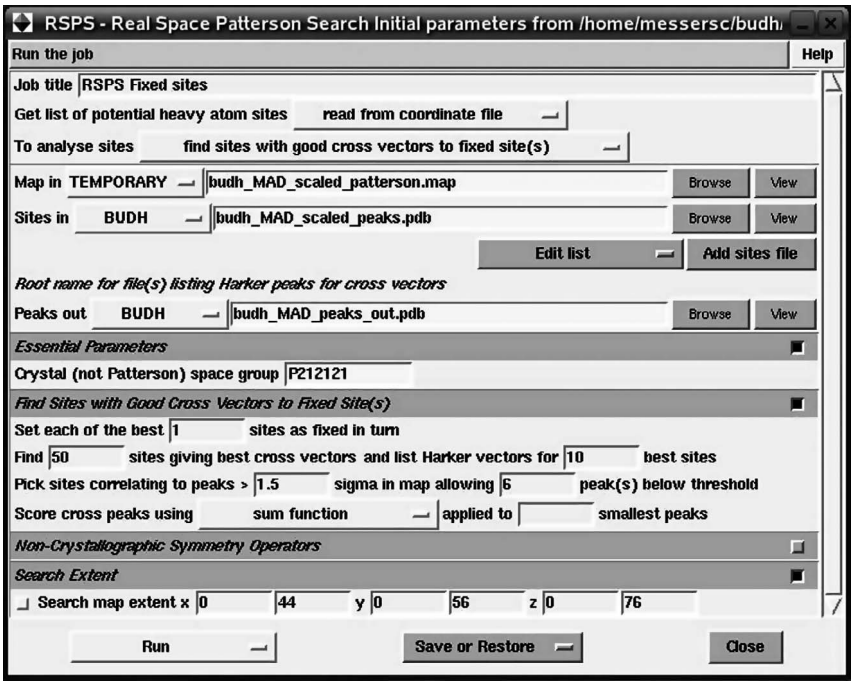


Fig. 11.14 CCP4i, RSPS parameters input for finding sites with good cross vectors to fixed site(s) for 4-BUDH (second run).

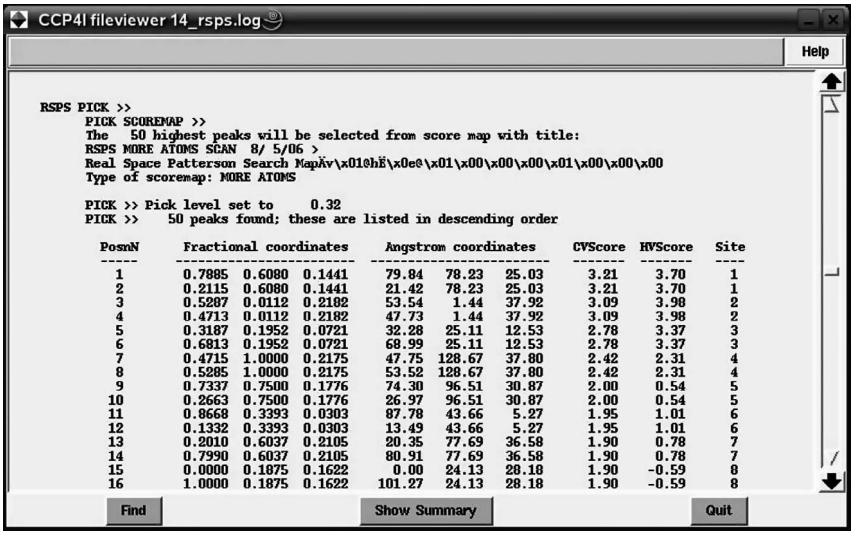


Fig. 11.15 Top of the peak list of the score map of MORE ATOMS SCAN of RSPS for 4-BUDH (second run). Number of fixed positions = 1. Position 1: 0.9995, 0.1649, 0.1809.

11.3
Comparison of the Results from SnB and RSPS

The coordinates of the sites from SnB together with the possible left-hand solution are given in the top part of Table 11.1. The RSPS sites are listed in the lower part of Table 11.1. The middle part of the table shows the coordinates of the SnB left-hand solution, together with their symmetry mates. The respective coordinate list for the SnB right-hand solution has not been given, as one can see that the RSPS solution corresponds to the SnB left-hand solution. The relevant sites are marked.

Both methods deliver equivalent correct solutions, but the direct methods reveal better results in cases with more heavy-atom sites to be determined.

Table 11.1 Comparison of the solutions from SnB and RSPS.

SnB solution, right hand			**SnB solution, left hand**		
x	***y***	***z***	***x***	***y***	***z***
0.999	0.666	0.821	0.001	0.334	0.179
0.524	0.512	0.784	0.476	0.488	0.216
0.788	0.608	0.642	0.212	0.392	0.358
0.674	0.697	0.929	0.326	0.303	0.071
Symmetry mates for SnB solution, left hand Symmetry operations: 1: x, y, z; 2: $-x+0.5, -y, z+0.5$; 3: $x+0.5, -y+0.5, -z$; 4: $-x, y+0.5, -z+0.5$					
1: 0.001	0.334	0.179	1: 0.476	0.488	0.216
2: 0.499	0.666	0.679 [a)] Site 1	2: 0.024	0.512	0.716 [a)] Site 5
3: 0.501	0.166	0.821	3: 0.976	0.012	0.784
4: 0.999	0.834	0.321	4: 0.524	0.988	0.284
1: 0.212	0.392	0.358	1: 0.326	0.303	0.071
2: 0.288	0.608	0.858	2: 0.174	0.679	0.571 [a)] Site 7
3: 0.712	0.108	0.642 [a)] Site 3	3: 0.826	0.197	0.929
4: 0.788	0.892	0.142	4: 0.674	0.803	0.429

RSPS solution

Site	***x***	***y***	***z***
1	0.999	0.165	0.181
3	0.215	0.609	0.143
5	0.527	0.013	0.217
7	0.678	0.192	0.070

a) The coordinates must each be added or subtracted by values of 0.5, respectively

References

Blessing, R. H., Smith, G. D., *J. Appl. Crystallogr.* **1999**, *32*, 664–670.
CCP4, *Acta Crystallogr.* **1994**, *D50*, 760–763.
Knight, S. D., *Acta Crystallogr.* **2000**, *D56*, 42–47.
Martins, B. M., Dobbek, H., Çinkaya, I., Buckel, W., Messerschmidt, A., *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 15645–15649.
Weeks, C. M., Miller, R., *J. Appl. Crystallogr.* **1999**, *32*, 120–124.

12
MIRAS and MAD Phasing with the Program SHARP

We assume that the heavy-atom positions or the positions of the anomalous scatterers have been determined by means of the methods used in Chapter 11. Several programs, which have been mentioned in Section 5.4.2, may be used to calculate the protein phases either from heavy-atom derivatives or anomalous diffraction data. Here, we will use the SHARP program (de La Fortelle and Bricogne, 1997) for our 4-BUDH example.

12.1
MAD Phasing with the Program SHARP for 4-BUDH

The SHARP program is operated by a web-based interface. The start screen (not shown) provides one with the possibility either to run SHARP in the auto-SHARP mode, or to start it from scratch or on the basis of a previous project, whereby SHARP offers two different projects for tutorial purposes. On starting SHARP, the Global Information Editor is opened (Fig. 12.1). Here, one must enter a project name, title, and the reflection data file. The file must be a multi-column MTZ file with the correct extension (*.data*.mtz), and be located in the *datafiles* directory. The chemical composition of the asymmetric unit must also be assigned. The actual values for 4-BUDH example are contained in Figure 12.1.

Next, we move to the Geometric Site Editor (Fig. 12.2), where one enters the coordinates of all heavy-atom or anomalous scatterer sites for the whole ensemble of compounds, crystals, and respective data sets. In the MIR case, one will have the native and several derivatives, denoted as compounds. Data sets will have been collected from these compounds, but they may also have been registered at different wavelengths and from distinct crystals. In the MAD case, one usually will have only one compound but data sets collected at different wavelengths, and they may also be from various crystals. The next step is the Compound Editor (Fig. 12.3). We have one compound only in the MAD experiment of 4-BUDH; hence, we must select all four Fe-sites from the global sites as C-sites for this compound. Additional compounds can be assigned by pressing the "New" button. Now, we can proceed to the Crystal Editor (Fig. 12.4), which holds information about the occupancies and temperature factors. Further crystals can

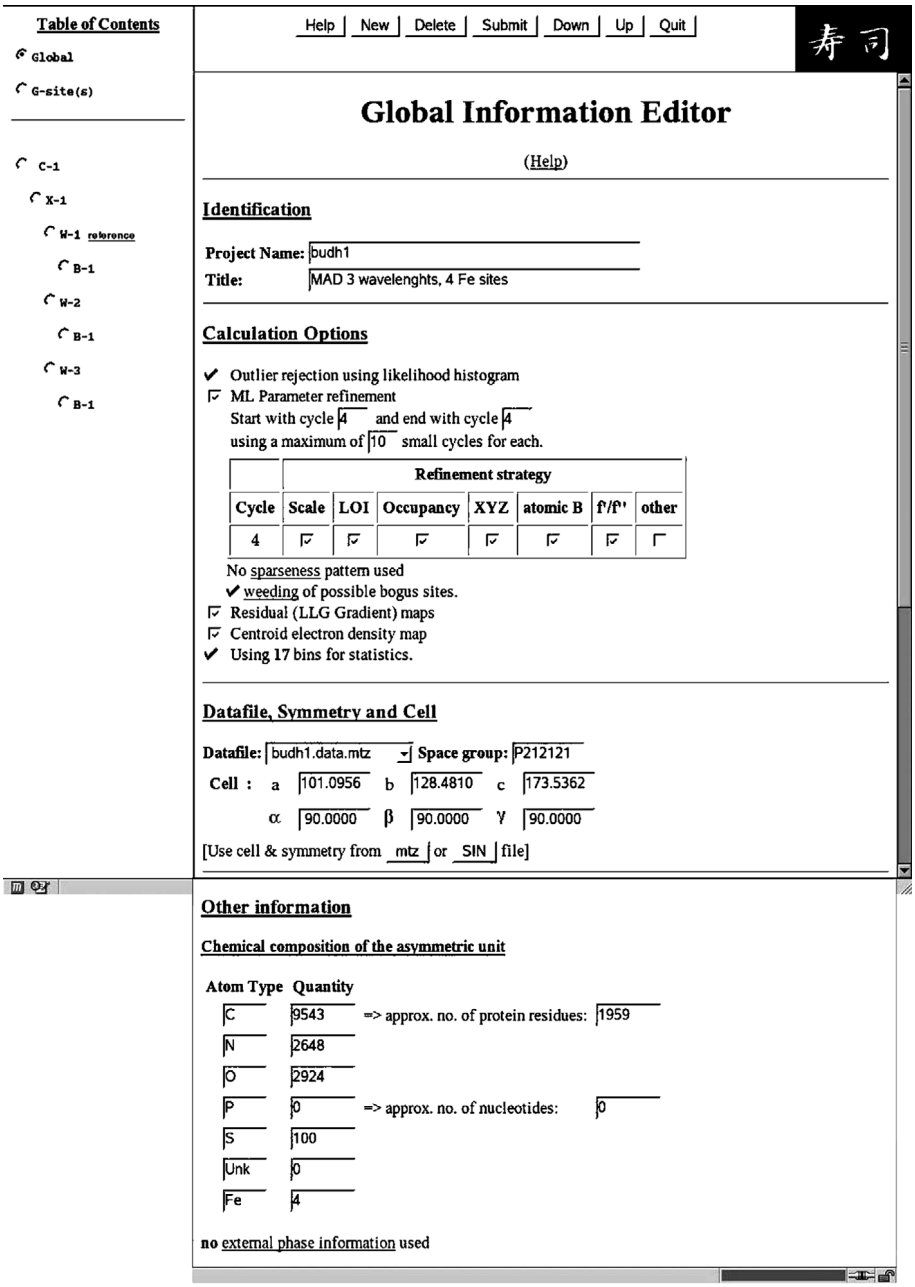


Fig. 12.1 SHARP Global Information Editor for 4-BUDH with 4 Fe-sites.

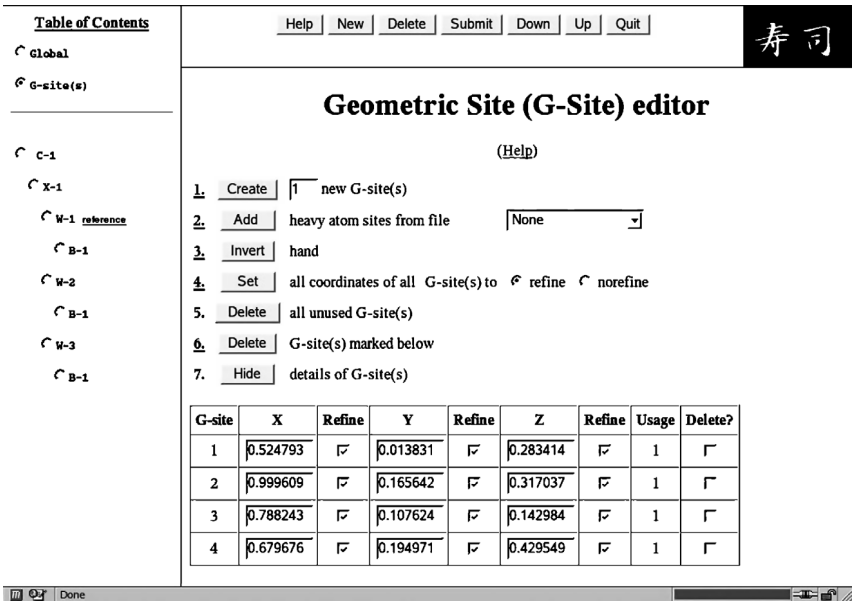


Fig. 12.2 SHARP Geometric Site Editor for 4-BUDH with 4 Fe-sites.

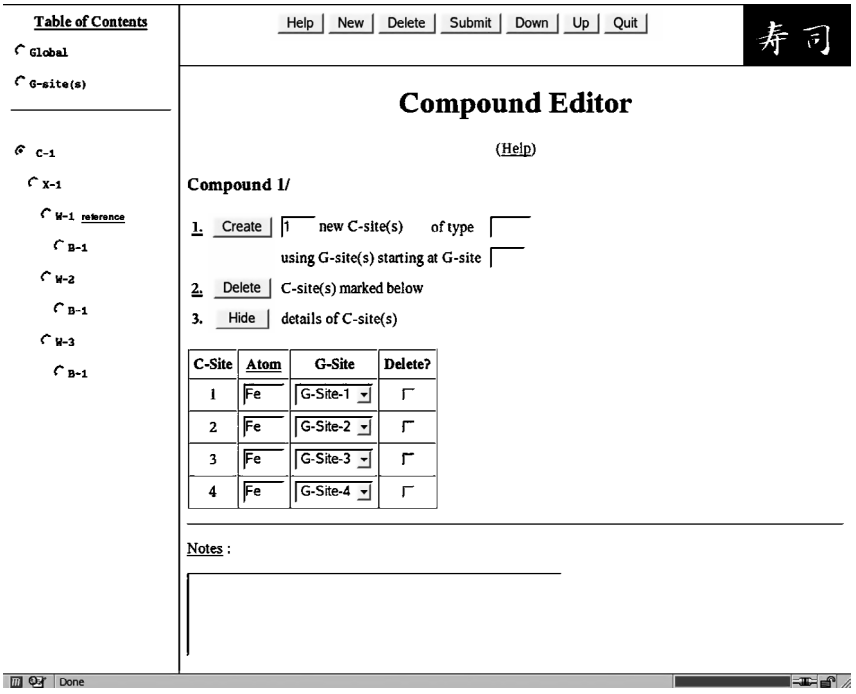


Fig. 12.3 SHARP Compound Editor for 4-BUDH with four Fe-sites.

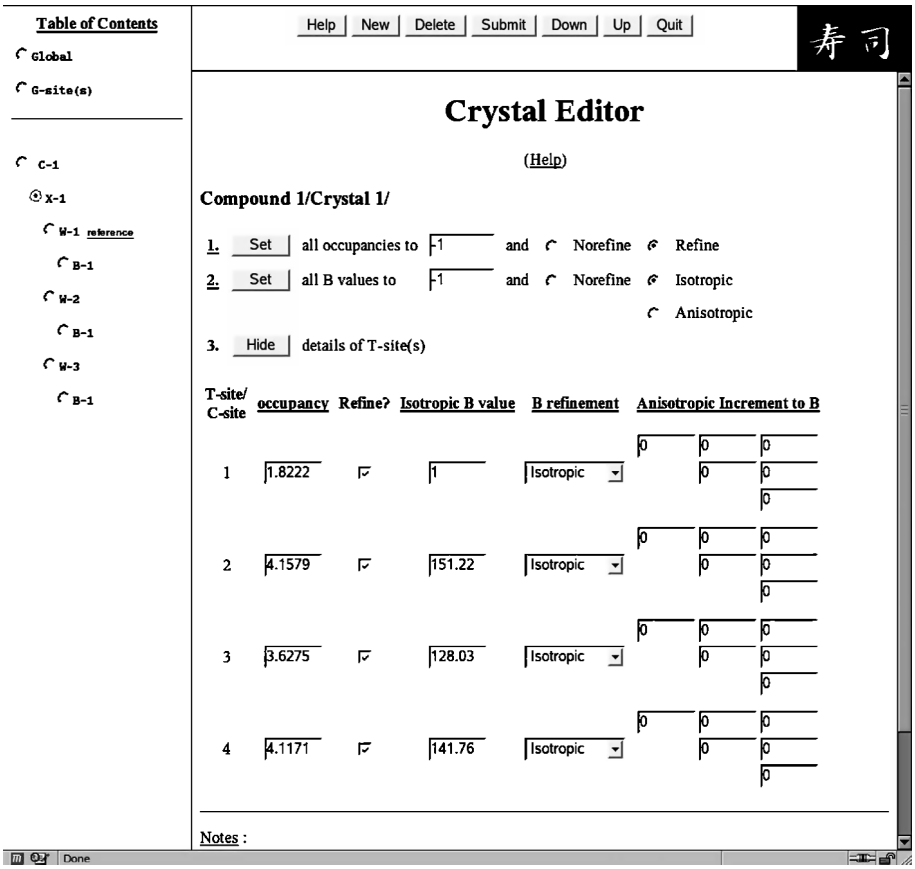


Fig. 12.4 SHARP Crystal Editor for 4-BUDH with four Fe-sites.

be added by pressing the "New" button. Different crystals of one compound should have the same sites, but they may have various occupancies and temperature factors.

As our 4-BUDH MAD-data have been collected at three different wavelengths, we have an individual Wavelength Editor for each wavelength. Figure 12.5 shows the respective editor for the remote wavelength. The lowest editor level is the Batch Editor. This is shown for the remote wavelength (Fig. 12.6). One must assign the correct columns for the Compound/Crystal/Wavelength/Batch and provide the relevant values of f′ and f″. For our example, the actual items for the other two wavelengths are: W-2 inflection, F_INF, SIGF_INF, DANO_INF, SIGDANO_INF f′ = –7.95, f″ = 2.83; W-3 peak, F_PK, SIGF_PK, DANO_PK, SIGDANO_PK, f′ = –6.84, f″ = 4.44.

If one has reached the last Batch Editor, then pressing the "Submit" button activates the running of the job. A listing of the input-file appears and can be

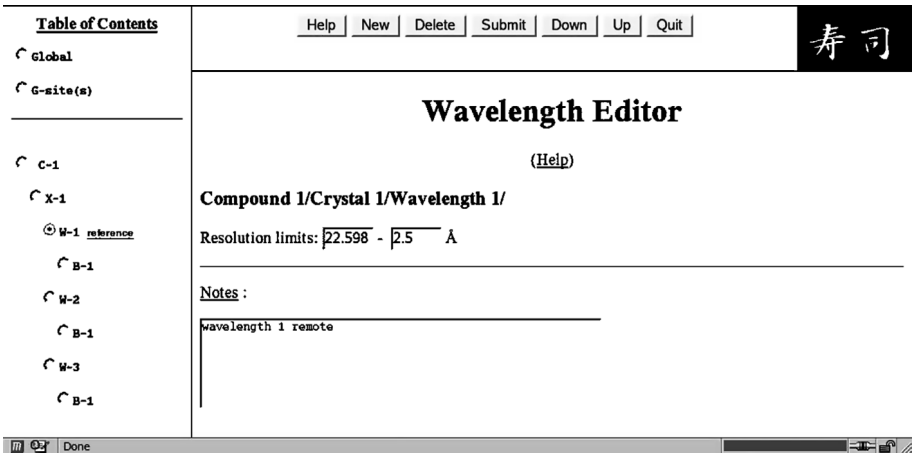


Fig. 12.5 SHARP Wavelength Editor for 4-BUDH with four Fe-sites.

checked for correctness before the actual job can be started. The results of the job are written to cardfiles, which may be used as input files for a new SHARP run, and to a central SHARP OUTPUT file. At the end of the logfile, residual maps can be viewed and analyzed to obtain the residual heavy-atom sites; the calculation of a centroid electron density map and phase improvement by density modification using the CCP4 program DM (Cowtan, 1994) can also be activated. We will not explain these output files in detail, but some important information has been extracted from these files (Table 12.1). As pointed out earlier, the determination of the heavy-atom or anomalous scatterer sites does not provide any information about the correct hand. Now, the usual way to do this involves calculating also the phases of the other hand (included in Table 12.1) and to examine some quality factors of the phase determination, density modification, or the quality of the electron density itself. These should be better for the correct hand, and the electron density should show right-handed α-helices if α-helices are present.

Until now, the Fe-sites in the Fe–S-clusters have not been resolved, and inspection of the residual maps showed that this is impossible with the quality of these phases. However, it is important to resolve the individual Fe-sites to obtain improved phases for a better experimental electron density. The 4-BUDH structure has one homotetramer per asymmetric unit, and the monomers within the tetramer are connected by NCS. Averaging of the MAD-phased electron density map will improve the quality of the map and allow resolution of the Fe-sites within each cluster. We will explain the procedure later, but here have used the resulting 16 Fe- plus 16 S-sites to recalculate the phases with SHARP. As the correct hand had been determined from the four Fe-sites phases, the phases were calculated only for the correct hand (Table 12.1).

The quality factors for all phase calculations are reasonable. Higher values for phasing power and FOM and lower values for R_{Cullis} correlate with better quali-

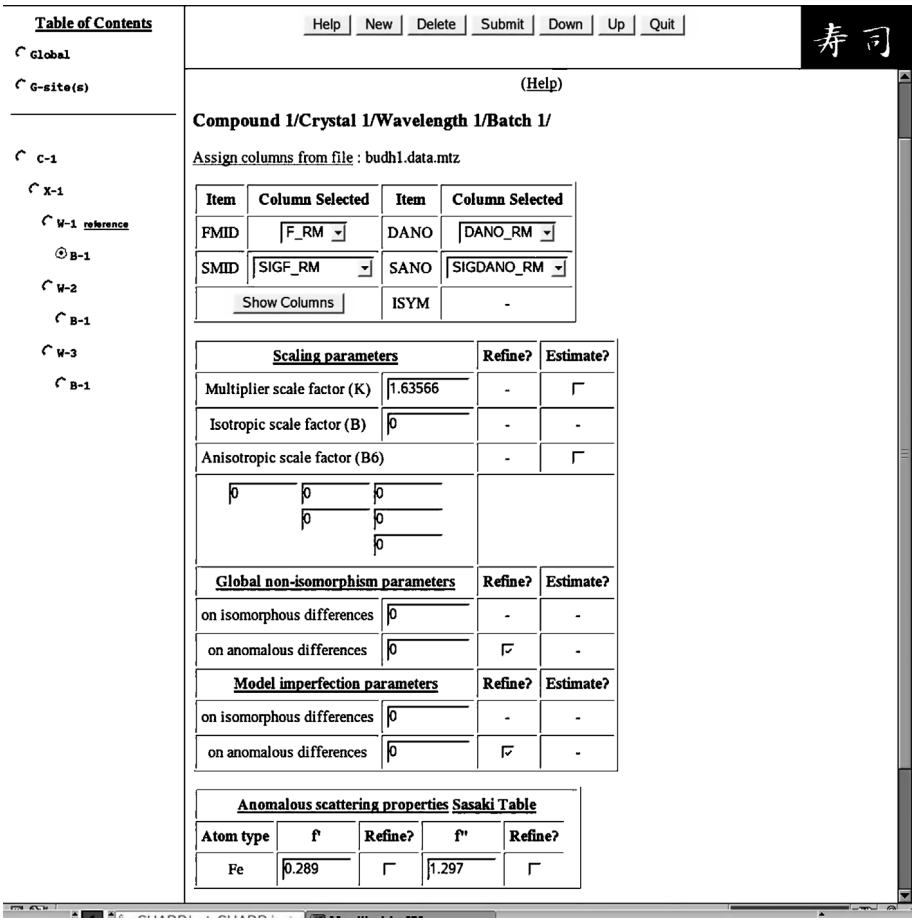


Fig. 12.6 SHARP Batch Editor for 4-BUDH with four Fe-sites.

ty of the electron density map, and should also be indicative of the correct hand. However, no differences can be found between the two possibilities in our 4-BUDH example (Table 12.1). Solvent flattening leads to a larger improvement of phases for the correct hand, and this is reflected in the respective quality factors. Lower values for SOLOMON and ICOEFL and higher values for "Overall correlation on |E|**2" are linked to a better quality of the electron density and to the correct hand. All relevant values for the right hand of 4-BUDH obey this requirement, but the differences are not very pronounced. A final decision can be made only by inspection of the corresponding electron density maps. This has been done and will be shown later. The analysis revealed the right hand as the correct solution. The phase calculation with 16 Fe-sites for the correct hand (Table 12.1) resulted in better values for all quality factors and, of course, in a better electron density map.

Table 12.1 Quality factors for MAD-phasing with SHARP and density modification with DM for 4-BUDH.

4 Fe-sites							**16 Fe-sites**		
	Right hand			**Left hand**			**Right hand**		
	ISO		**ANO**	**ISO**		**ANO**	**ISO**		**ANO**
	acent	**cent**		**acent**	**cent**		**acent**	**cent**	
Phasing power[a)]	0.887	0.802	0.977	0.887	0.802	0.997	1.016	0.807	1.061
R_{Cullis}[b)]	0.769	0.820	0.832	0.769	0.820	0.832	0.747	0.792	0.808
FOM[c)]	0.507	0.310		0.507	0.310		0.539	0.339	
Solvent flattening with DM									
SOLOMON[d)]	0.374			0.391			0.349		
ICOEFL[e)]	0.334			0.335			0.320		
Overall correlation on |E|**2[f)]	0.554			0.542			0.610		

a) Phasing Power = $\langle|F_{h\,calc}|/[\text{phase-integrated lack of closure}]\rangle$.
b) $R_{Cullis} = \langle\text{phase-integrated lack of closure}\rangle/\langle|F_{PH} - F_P|\rangle$.
c) FOM = Figure of merit, as given in Eq. (5.68).
d) SOLOMON = sd of solvent before flattening/sd of protein.
e) ICOEFL = Overall R-factor RO, a simple R-factor between structure factor amplitudes from the modified map and the observed data.
f) Overall correlation on |E|**2 = Overall correlation on |E|**2 between structure factor amplitudes of the observed data and the modified map.

References

Cowtan, K., *Joint CCP4 and ESF-EACMB Newsletter on Protein Crystallography* **1994**, *31*, 34–38.

De La Fortelle, E., Bricogne, D., *Methods Enzymol.* **1997**, *276*, 472–494.

13
Molecular Replacement

The method of molecular replacement was discussed in Section 5.5, together with the commonly used programs. Recently, activities have been completed to automate the phase determination by molecular replacement methods. One possibility is to use the program MrBUMP, which is a project of the CCP4 consortium available under the following web-address: http://www.ccp4.ac.uk/MrBump. Another option is CaspR (Claude et al., 2004). Both approaches use homology modeling for the construction of a powerful structural start model. It should be noted that the use of such a web server is not advisable if your project is confidential.

Here, we will not discuss a very complicated example, because this would go beyond the scope of this practical part of the book. However, the example does feature several points of view, which may lead to failure of the procedure if not correctly addressed.

13.1
Phase Determination of PKC-iota with Program Molrep

We use the CCP4-supported program Molrep (Vagin and Teplyakov, 1997) for the performance of the phase determination by molecular replacement. The subsequent procedure has been used to solve the crystal structure of the catalytic domain of human PKC-iota (Messerschmidt et al., 2005). The overall structure of the catalytic domain of all protein kinases is very similar, and as several crystal structures of such domains have been solved they can be used as structural search models for the structure determination of related catalytic kinase domains. In the case of human PKC-iota, the crystal structure of the catalytic domain of PKC-theta, a member of another PKC subfamily, had been solved and was available under the Protein Data Bank code 1XJD (Xu et al., 2004). The catalytic subunit of protein kinases consists of an N- and C-terminal lobe (Fig. 13.1). The C-terminal lobes superimpose very well for all kinase domains, but the N-terminal lobes may adopt rather different positions with respect to the C-terminal lobe. This must be considered when constructing the search model. Initially, an amino acid sequence alignment between PKC-iota and PKC-theta was made, and all parts that did not align well were omitted from the

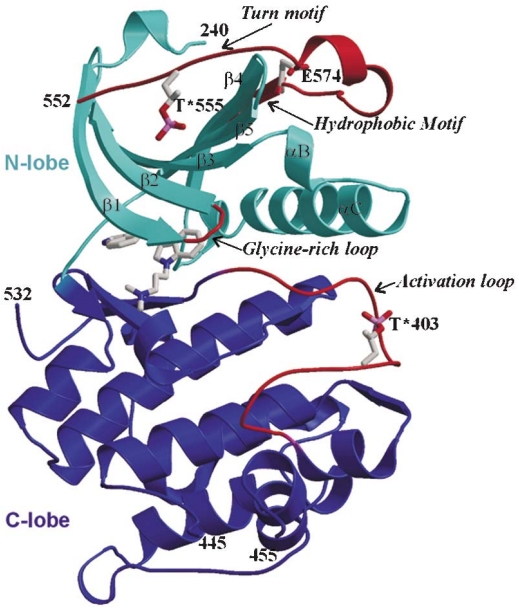


Fig. 13.1 Overall structure of the PKC-iota-BIM1 complex. (Reproduced by permission of Elsevier, from Messerschmidt et al., 2005.)

search model. In the remaining model, all non-identical residues were mutated to alanine, and only the C-terminal lobe (residues 440–649 in the PKC-theta structure) was used.

The interface for initial parameter input for Molrep is shown in Figure 13.2. One needs the reflection input MTZ-file and the coordinate file of the search model only. The number of molecules per asymmetric unit had been determined as 1 by means of the Matthews parameter. The molecular replacement method should work in straightforward manner if about 50% or more of the scattering power is known and only one molecule has to be searched for in the asymmetric unit. The rotation function should be sharper because only one rotation has to be determined. The translation function will also be easier to interpret because only one set of cross-vectors has to be identified. Furthermore, the weights of the corresponding vector sets are higher when one molecule is to be searched for compared to more molecules.

In Section 10.3 we explained the space group determination for PKC-iota with space groups $P3_12(1)$ or $P3_22(1)$ as remaining possibilities. Now, we must run the molecular replacement with both space group symmetries. Figure 13.3 displays the crucial part of the Molrep output for space group $P3_22(1)$. The highest peak of the rotation function has an RF/sigma of 4.95 and rotational angles of alpha = 29.57, beta = 35.29 and gamma = 212.91. Translation functions are then performed with each of the top ten solutions of the rotation function. The results of the translation function for only the first two rotation function maxima are shown. The best solution is for the highest rotation function maximum with a correlation factor of 0.387 and translations of Xfrac = 0.525, Vfrac = 0.464, and

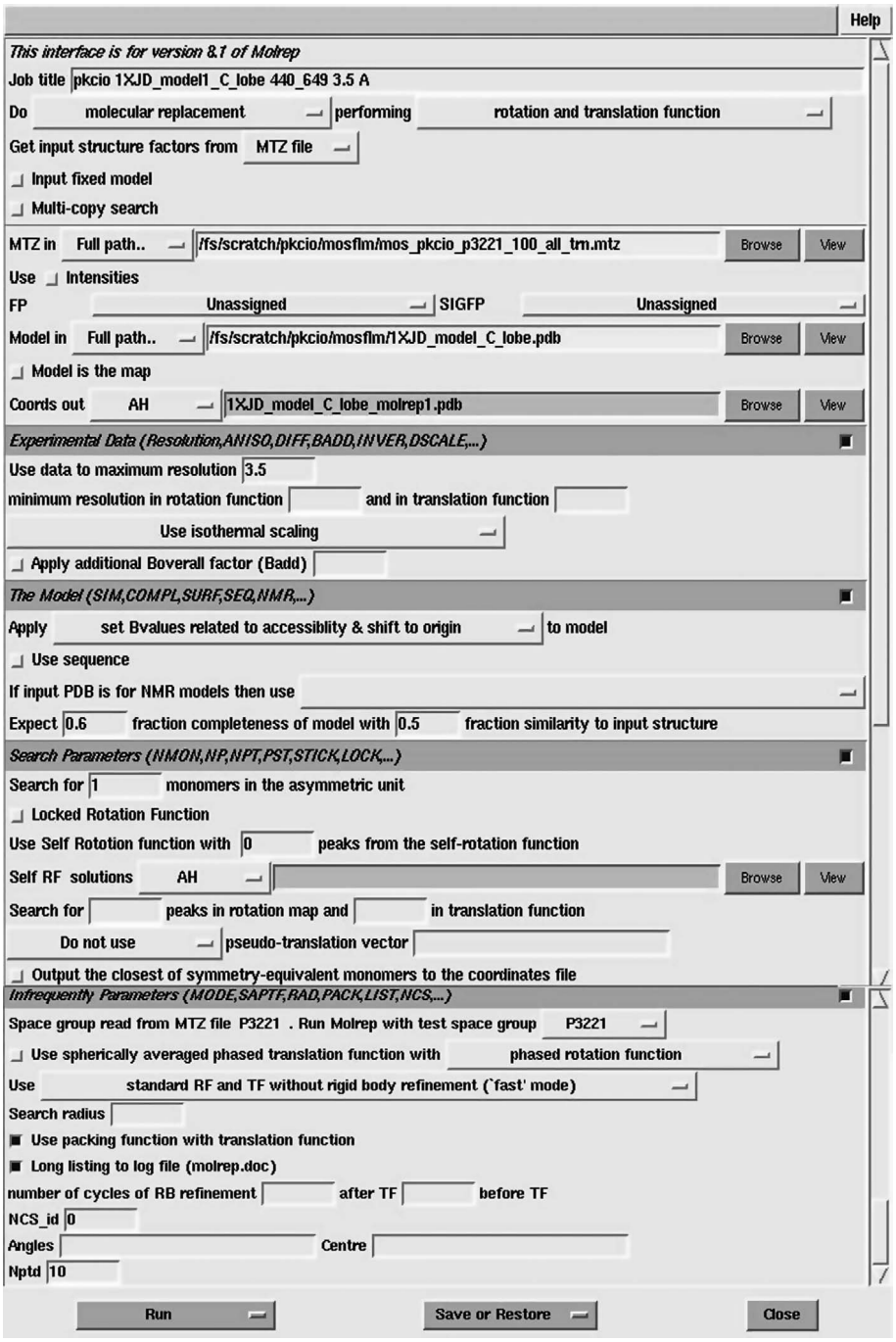


Fig. 13.2 CCP4i, Molrep, Initial Parameters for Molecular Replacement.

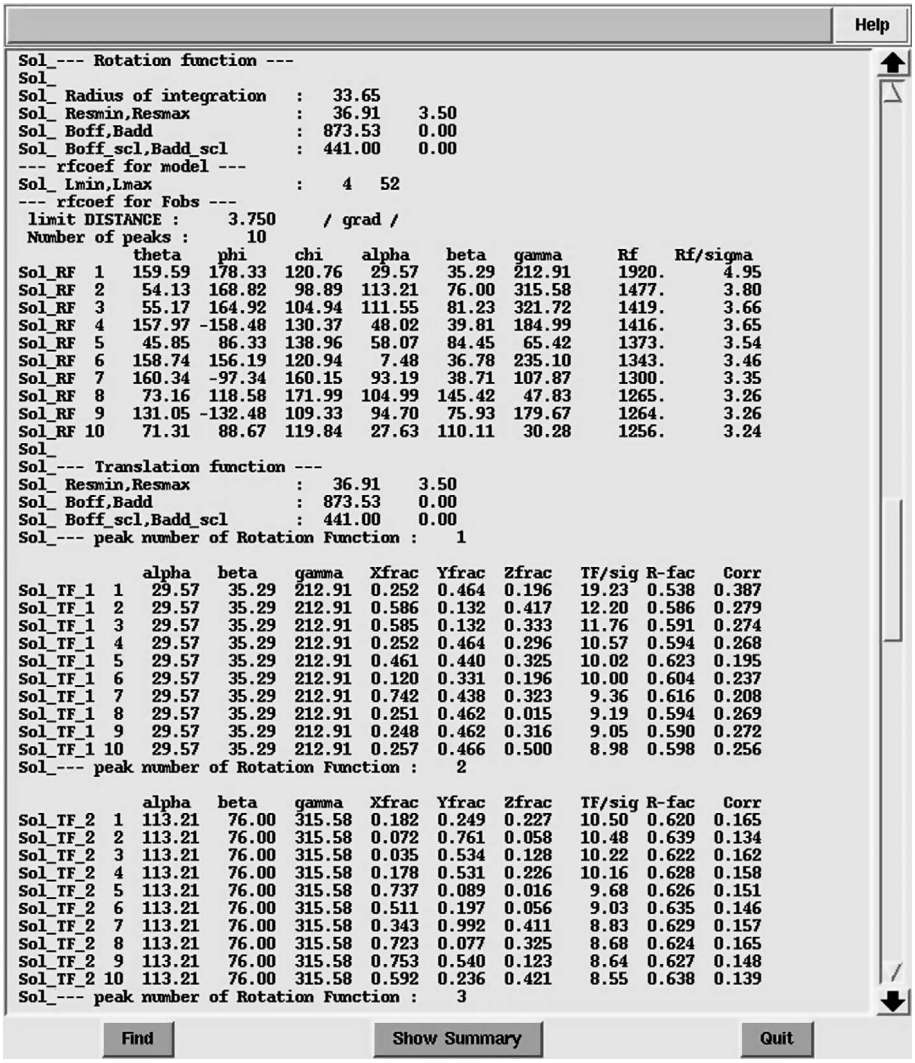


Fig. 13.3 CCP4i, Part of Molrep output for PKC-iota, space group $P3_22(1)$.

Zfrac = 0.196. The corresponding listing for space group $P3_12(1)$ is depicted in Figure 13.4. The rotation function is identical of course because it depends only on the point group symmetry. The highest correlation factor of the translation functions is 0.245, which is much lower than that for space group $P3_22(1)$. The program Molrep writes the transformed coordinates to an output file in PDB-format. It can be used directly for phase and subsequent electron density map calculation. The electron density map for PKC-iota computed with the best solution from space group $P3_22(1)$ was of good quality and contained the electron

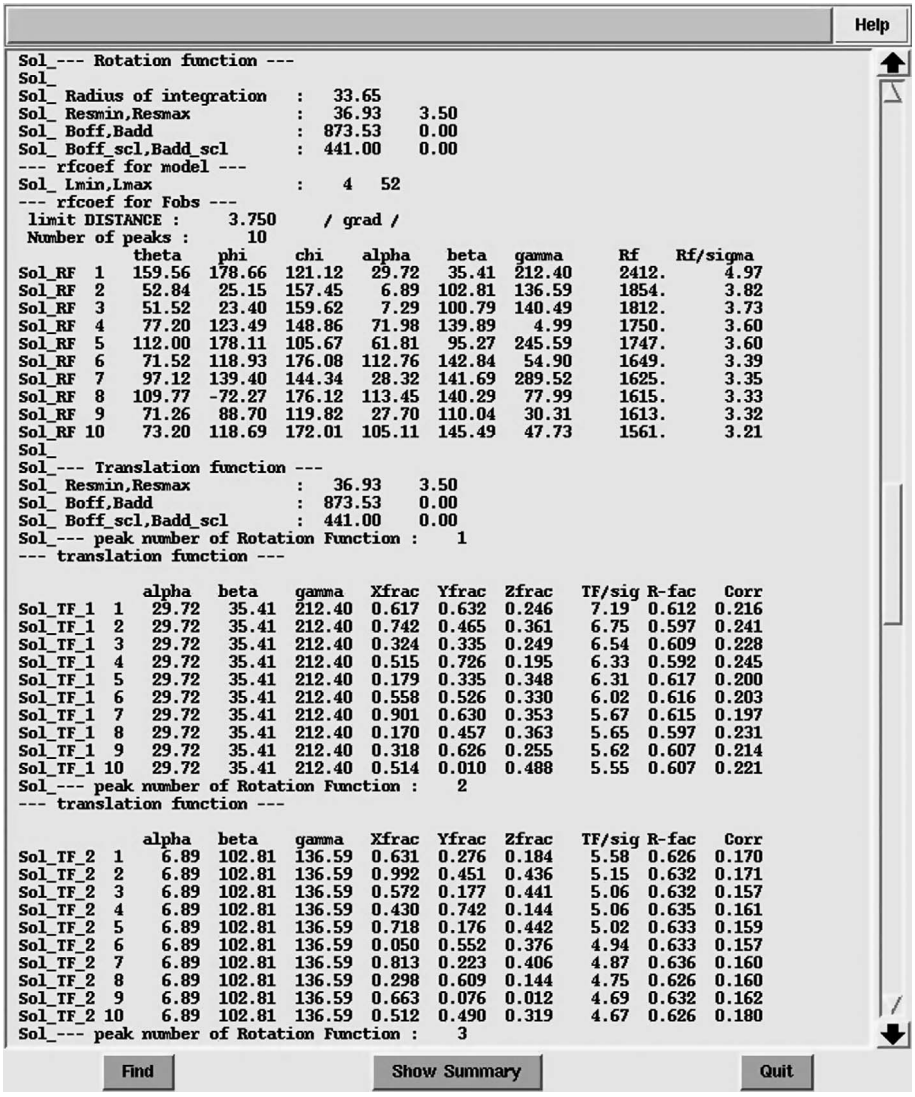


Fig. 13.4 CCP4i, Part of Molrep output for PKC-iota, space group $P3_12(1)$.

density for the missing N-terminal lobe. The complete model could be developed in several model building and structure refinement cycles (for further details, see Messerschmidt et al., 2005).

References

Claude, J.-B., Suhre, K., Notredame, C., Claverie, J.-M., Abergel, C., *Nucleic Acids Res.* **2004**, *32*, W606–W609.

Messerschmidt, A., Macieira, S., Velarde, M., Bädeker, M., Benda, C., Jestel, A., Brandstetter, H., Neuefeind, T., Blaesse, M., *J. Mol. Biol.* **2005**, *352*, 918–931.

Vagin, A.A., Teplyakov, A., *J. Appl. Crystallogr.* **1997**, *30*, 1022–1025.

Xu, Z.-X., Chaudhary, D., Olland, S., Woldrom, S., Czerwinski, R., Malakian, K., et al. *J. Biol. Chem.* **2004**, *279*, 50401–50409.

14
Averaging about Non-Crystallographic Symmetry (NCS) for 4-BUDH

The theoretical basis of NCS electron density averaging was explained in Section 6.3.

The first NCS averaging programs, which were written by Bricogne in 1976, formed the basis for the relevant routines which are available in program packages today. Program systems such as PHASES (Furey and Swaminathan, 1997), MAIN (Turk, 1995), DM/DMMULTI (Cowtan, 1994), which is a part of the CCP4 program suite, and the Uppsala Software Factory (USF) (Kleywegt and Jones, 1994) contain respective routines.

The NCS or local symmetry may be proper (also termed closed). In this case, the local mask will cover the whole complex, the whole homo-tetramer in our 4-BUDH example. This may be very useful at the start of the procedure when the boundaries of the whole complex can often be determined easily, but not the envelope of the monomer. If the local symmetry is open or improper, the envelope can cover only the monomer.

The averaging procedure consists of determining the NCS symmetries, defining a molecular envelope or mask, averaging of the electron density, and reconstituting the averaged map for Fourier back-transforming. This back-transformation corresponds to a solvent-flattening step because the space outside the molecular masks has been set to zero. Such a map should have an improved quality and can be used for further cycles of averaging or averaging plus phase extension. In many circumstances, it is sufficient simply to average the electron density map and to build an improved structural model into this averaged map. The atomic model for the whole asymmetric unit is then generated with the aid of the current NCS operators, and a new crystallographic structure refinement can be started. The refined structural model is used to improve the NCS operators, which are then used when the relevant electron density map is averaged.

Here, we will demonstrate NCS averaging at the example of 4-BUDH using the relevant programs of the USF throughout. These programs are compatible with the CCP4 program suite and the modeling program "O" (Jones et al., 1991), which has been mainly employed by the present author and will be discussed briefly below. This was shown to work reliably and smoothly. The present example does not include cyclic averaging or phase extension, which are subjects for more specialized tracts.

14.1
Determination of NCS Operators for 4-BUDH

The centers of the four [4Fe–4S] clusters in one homotetramer in the unit cell of 4-BUDH as determined by SnB are shown in Figure 14.1. The homotetramer exhibits a proper 222 symmetry. The local diads are drawn into Figure 14.1. The local symmetry axes can be determined from a Patterson self-rotation function, but this rotation function will reveal the character of such axes (e.g., twofold, threefold, etc., or general rotation) and their orientation in space without position of translation only. A self-rotation function for 4-BUDH detected three twofold local axes perpendicular to each other, but we will not discuss this analysis because the calculation of such a rotation function by, for example, the program GLRF is simple and the interpretation of the results straightforward.

The NCS operators can be easily determined if heavy atoms have equally bound to the NCS monomers or anomalous scatterers such as the Fe–S-clusters are present. Unfortunately, one needs at least three nonlinear heavy atoms or

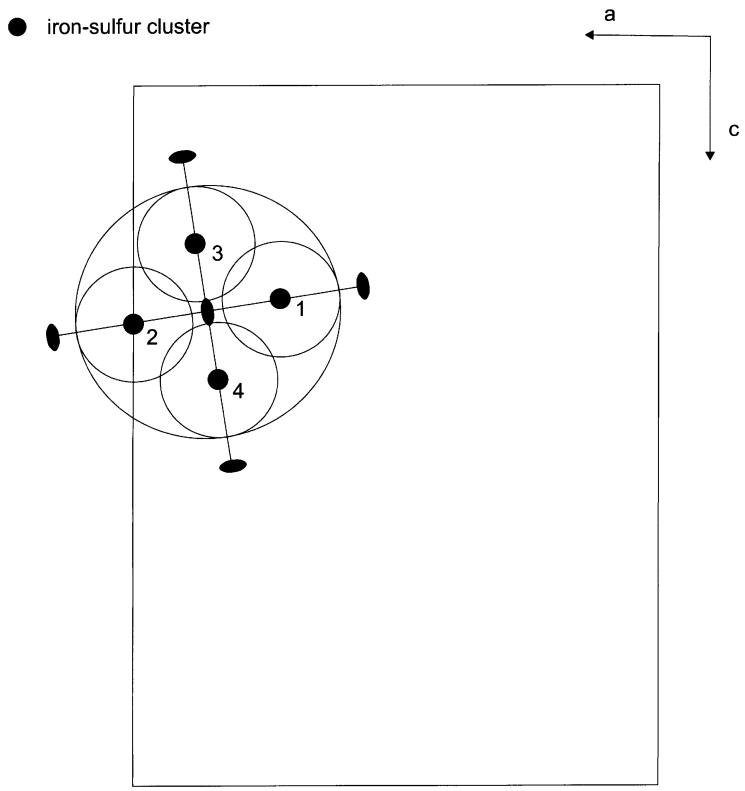


Fig. 14.1 The position of the homotetramer plus the centers of Fe–S-clusters in the unit cell of 4-BUDH.

anomalous scatterers per monomer in order to determine unambiguously the NCS operators directly from their positions. Although this is not the case in the present example, we do have an experimental electron density map which can be used for this purpose. This map has been phased with the program SHARP to a resolution of 2.5 Å using one Fe-site per [4Fe–4S]-cluster, four Fe-sites per asymmetric unit, and subsequently subjected to phase improvement by a solvent-flattening optimization run with the program DM.

The NCS operators are determined using the USF program NCS6D, which requires a degree of preparation before it is run. The electron density map to be searched must cover a sufficiently large area, and this has been done by performing the CCP4 routine EXTEND with new map dimensions in the range of –0.5 to 1.2 for *x*, *y*, *z*, each (fractional coordinates). The name of the actual file was "eden_flat_43.3pc_rh_ext.ccp4". Furthermore, we need a skeletonized part of the map around the Fe-site number 1. Such a skeletonization can be prepared with the USF program MAPMAN. The relevant skeletonization routine is based on the Greer algorithm (Greer, 1974). All USF programs can be run not

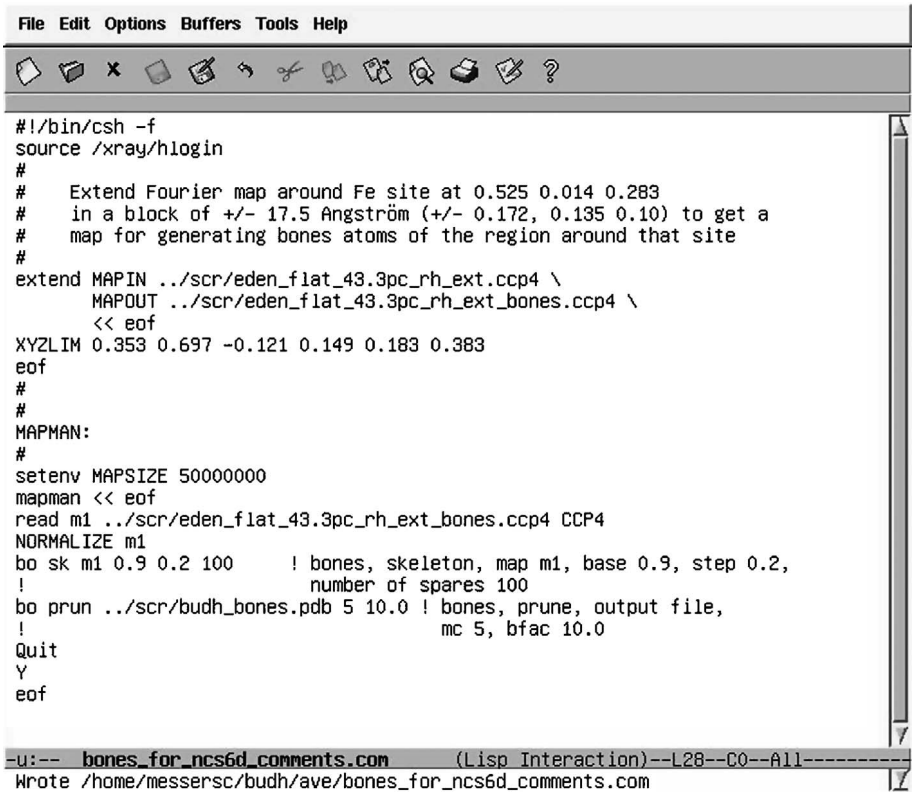


Fig. 14.2 Input file for the generation of bones atoms with USF program MAPMAN for 4-BUDH.

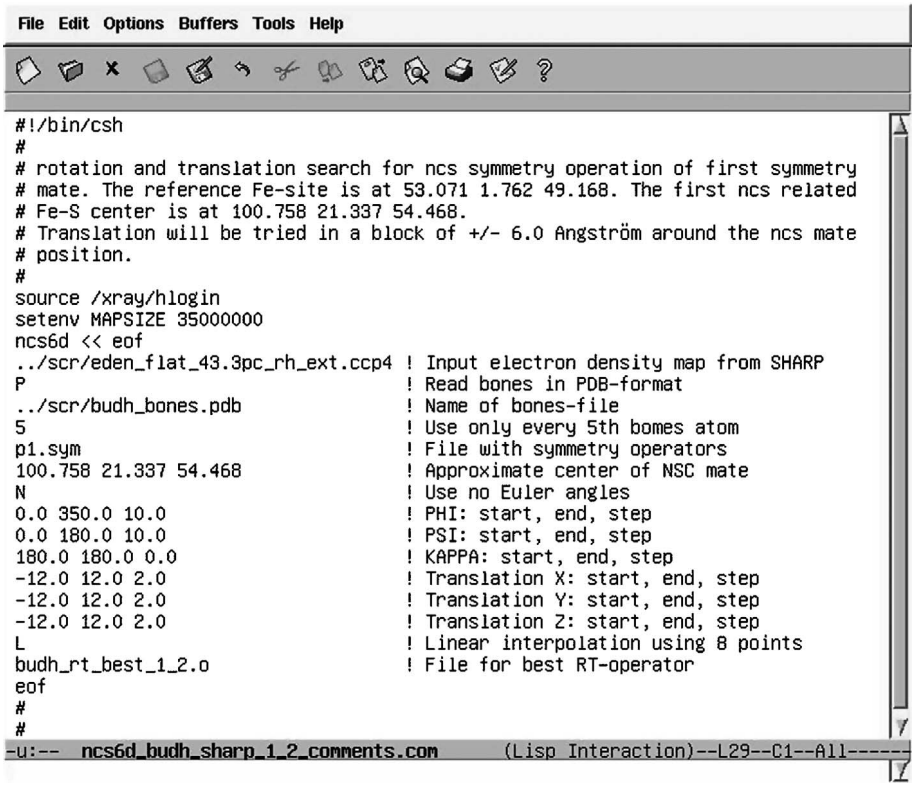


Fig. 14.3 Input file for NCS operator search with USF program NCS6D for mapping monomer 1 onto monomer 2 for 4-BUDH.

only interactively but also in batch mode. We present here input files that can be used for the batch mode. Figure 14.2 shows the input necessary to generate the bones file. First, a block of ±17.5 Å around the Fe-site number 1 is separated from the input map to obtain a map for generating bones atoms of the region around that site (CCP4 EXTEND step). Then, MAPMAN is invoked to run the actual skeletonization (bones skeleton map m1), storing the bones atoms in PDB-format in the file "budh_bones.pdb". Figure 14.3 displays the input for NCS6D. The program takes the bones atoms around Fe-site number 1 and performs a rotational/translational search around the NCS symmetry mate Fe-site number 2. Input and output files and parameters have been explained in Figure 14.3. The translation will be tried in a block of ±12.0 Å around the NCS mate position. Similar input files have been created for the mapping of monomer 1 onto monomer 3 and monomer 4, respectively.

The top of the output file with the best 100 solutions for mapping monomer 1 onto monomer 2 is shown in Figure 14.4. The correct solution had a correla-

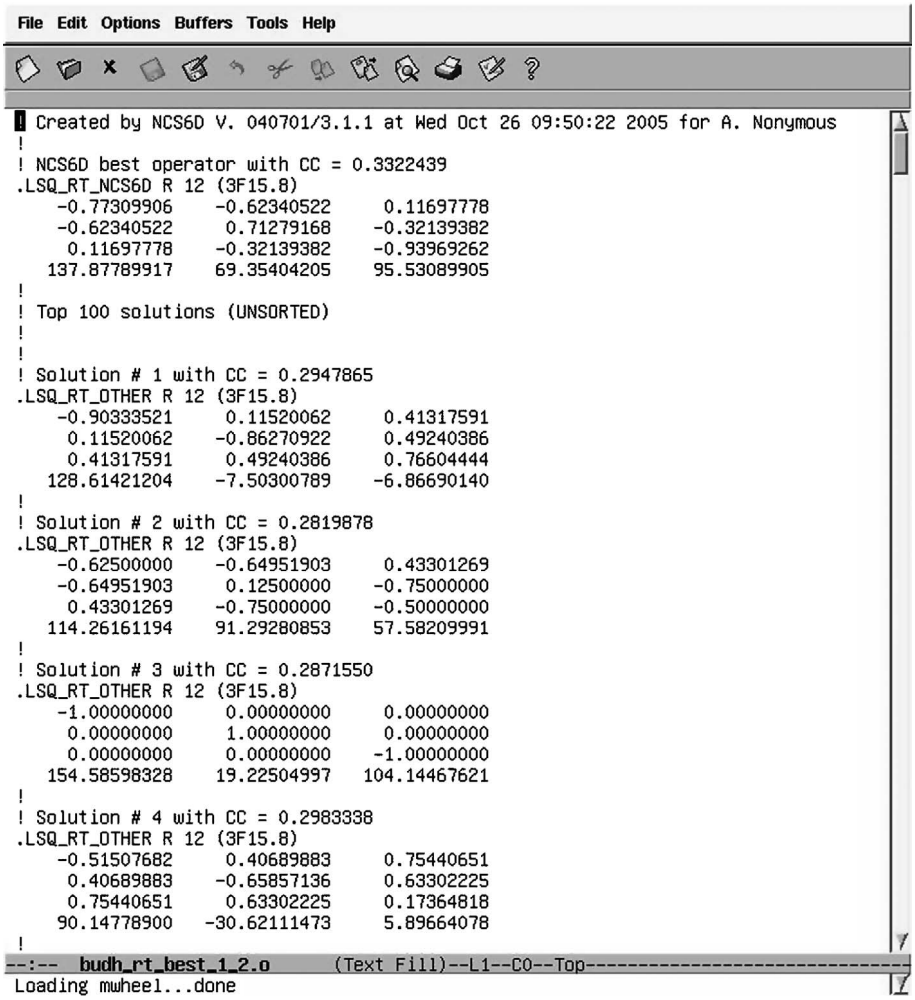


Fig. 14.4 Top of the output file of NCS6D with the best 100 solutions mapping monomer 1 onto monomer 2.

tion coefficient of 0.332 and rank 1. The corresponding values for the two other NCS operators are 0.341; rank 1 and 0.391; rank 4. Rank 4 for the third NCS operator was not optimal, but the correlation coefficient was close to the best value (0.407) and therefore still a candidate for checking. The USF averaging programs perform all coordinate transformations in a Cartesian coordinate system. Atom positions in fractional coordinates must be converted into Cartesian coordinates. This and the reverse transformation can be done in the USF program MOLEMAN (options FRACtional_to_cartesian and CARTesian_to_fractional). The orthogonalization conventions are consistent with those used in the Protein Data Bank. The NCS operator consists of a 3×3 rotation matrix and a

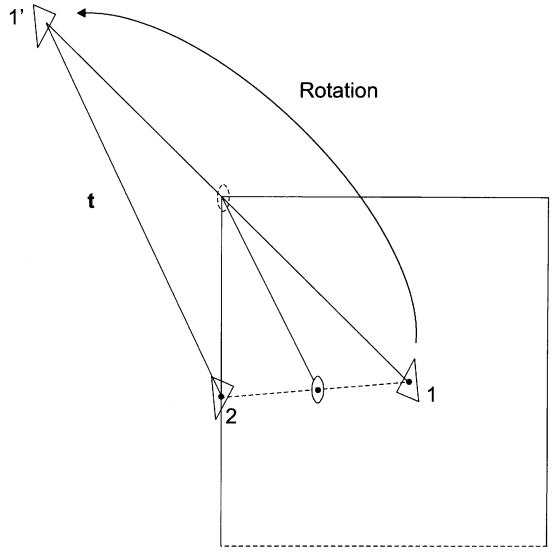


Fig. 14.5 Explanation of the action of a rotation/translation operator.

File Edit Options Buffers Tools Help

```
#!/bin/csh
source /xray/hlogin
setenv MASKSIZE 10000000
mama -b << END-mama
      new grid 120  156  208
      new cell  101.0960  128.4810  173.5360   90.0000   90.0000   90.0000
      new radius 4.0
      new pad 10 10 10
      new ball m1 53.071 1.762 49.168 20.0
      fill m1
      nbr m1
      isl m1
      smooth m1 10
      ov sym p1.sym (get the 0 spacegroup file)
      fi m1
      co m1
      co m1
      is m1
      ex m1
      ex m1
      ov tri m1 4.5
      write m1 mask_sharp.mask
      quit
END-mama
```

```
-u:--  mask_sharp_bw.com      (Lisp Interaction)--L25--C0--All----------------------
Wrote /home/messersc/budh/ave/mask_sharp_bw.com
```

Fig. 14.6 Input file for mask generation with USF program MAMA for 4-BUDH.

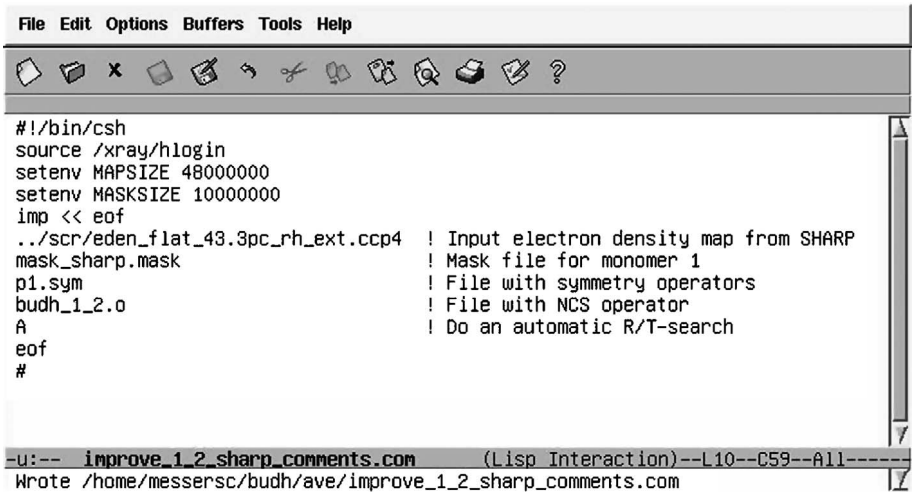


Fig. 14.7 Input file for the improvement of the NCS operator between monomer 1 and monomer 2 with USF program IMPROVE for 4-BUDH.

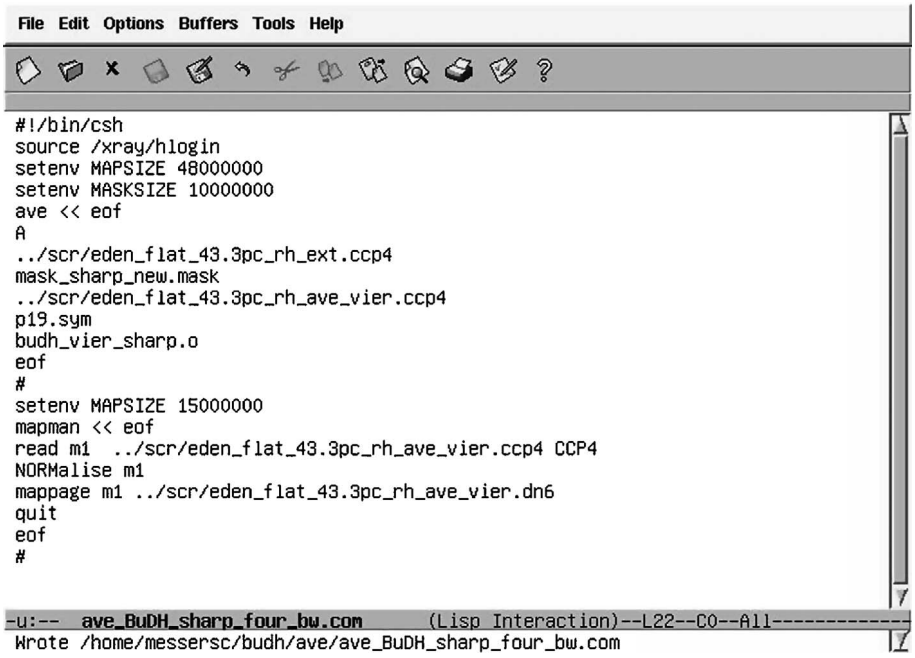


Fig. 14.8 Input file for electron density averaging with program AVE for 4-BUDH.

3×1 translation vector. The action of the NCS operator for mapping monomer 1 onto monomer 2 is illustrated in Figure 14.5. First, the coordinates of monomer 1 are rotated by a rotation located at the origin to position 1′ and then shifted by translation **t** to the end position 2.

Next, it is very useful to run the USF program IMPROVE to check the correctness of the NCS operator and to improve its position, which is very important for the subsequent averaging procedure. For this, one must generate a mask file; this can be done with the USF program MAMA. At present, we have minimal information about the extension of an individual monomer. The 4-BUDH monomer has a molecular mass of about 50 kDa, and a sphere around the Fe–S-cluster with a radius of about 20 Å is a reasonable conservative first approximation. The respective input file is shown in Figure 14.6. The chosen grid and unit cell parameters must be consistent with those of the used electron density map. The input for program MAMA is depicted in Figure 14.7 and explained by the relevant comments. The IMPROVE procedure enhanced the correlation values as follows: monomer 1 to monomer 2: from 0.057 to 0.197; monomer 1 to monomer 3: from 0.170 to 0.189; monomer 1 to monomer 4: 0.061 to 0.185.

14.2
Electron Density Map Averaging for 4-BUDH

The actual electron density averaging is done with the USF program AVE, and the respective input file is shown in Figure 14.8. The input is very easy to perform. Parameter "A" stands for mode averaging, after which the names of the electron density map to be averaged and the mask file must be given. Next, the name of the averaged output map is entered. The symmetry operators for the space group $P2_12_12_1$ (Nr. 19) (file p19.sym) and the NCS operators including the identity (file budh_vier_sharp.o) must be given. The logfile of AVE contains as essential parameters the correlation factors for the individual NCS operations, which should be greater than 0.15 at the initial stages, which is the case for our 4-BUDH example, and better than 0.40 during later stages as a rule of thumb. Subsequently, MAPMAN is run to normalize the averaged electron density map and to convert it from a CCP4 style to a DSN6 style file, which can be read into the graphics program system "O". We have now reached a point, where one should inspect the electron density maps and start the model building. We will use the graphics modeling program "O" and make some related introductory notes in Chapter 15.

References

Bricogne, G. *Acta Crystallogr.* **1976**, *A32*, 544–548.

Cowtan, K., *Joint CCP4 and ESF-EACMB Newsletter on Protein Crystallography* **1994**, *31*, 34–38.

Furey, W., Swaminathan, S., *Methods Enzymol.* **1997**, *277*, 590–620.

Greer, J., *J. Mol. Biol.* **1974**, *82*, 279–301.

Jones, T.A., Zou, J.Y., Cowan, S.W., Kjelgaard, M., *Acta Crystallogr.* **1991**, *A47*, 110–119.

Kleywegt, G.J., Jones, T.A., Halloween … Masks and Bones. In: Bailey, S., Hubbard, R., Waller, R. (Eds.), *From First Map to Final Model*, pp. 59–66. Daresbury Laboratory, Warrington, **1994**.

Turk, D., *American Crystallography Association Annual Meeting* **1995**, *27* (ISSN 0596-4221) p. 54, Abstract 2m.6.B.

15
Model Building and More

15.1
A Very Personal Short Introduction to the Computer Graphics Modeling Program "O"

The computer graphics modeling program "O" (Jones et al., 1991) is the present author's favorite model-building program. The system is very versatile, allows a personal tailoring, shares a common environment with the USF programs, and is compatible with the CCP4 program suite. Here, we will not present a detailed instruction manual because this has been provided on A. Jones' "O" webpage (http://alpha2.bmc.uu.se/alwyn/) or the tutorial from G. Kleywegt, "O" for morons, accessible under http://yray.bmc.se/usf/. The latest release of "O" is now version 10, and the program can be obtained from A. Jones on request for several common computer platforms. It contains important changes in the *Decor, Sprout, SST* systems. The Decor commands are used during the map interpretation stage of model building and have been reviewed in a detailed overview (Jones, 2004). The skeleton (a skeletonized electron density generated, for example, in the program MAPMAN) is useful to sketch how a macromolecule folds in space. The program TRACE provides a more detailed representation of the molecule, and has local directionality. It can be built more or less automatically with the *Sprout* commands, or interactively with the *SST* commands. The step in going from the TRACE to a molecule (or part of a molecule) with the sequence assigned is assisted by the *Decor* commands. These require that the user has built a part of the TRACE molecule, without any gaps, and has created a continuous skeleton that follows the TRACE. With the tutorials and manuals in hand, the novice should be able to become familiar with the system in a rather straightforward manner.

Here, we present a method for model building with "O", which is rather simple minded and little automated, requires minimal reading of the manuals, and is also rather quick. First of all, the program "O" must be installed on your computer graphics system. The advanced user can make this directly, but the "normal" user should ask his or her computer administrator to do this. Running "O" for the first time needs certain preparations, since after starting "O" one is asked for an existing "O" data base. The "O" data base holds all information related to the model building process. Clearly, as one does not have such a

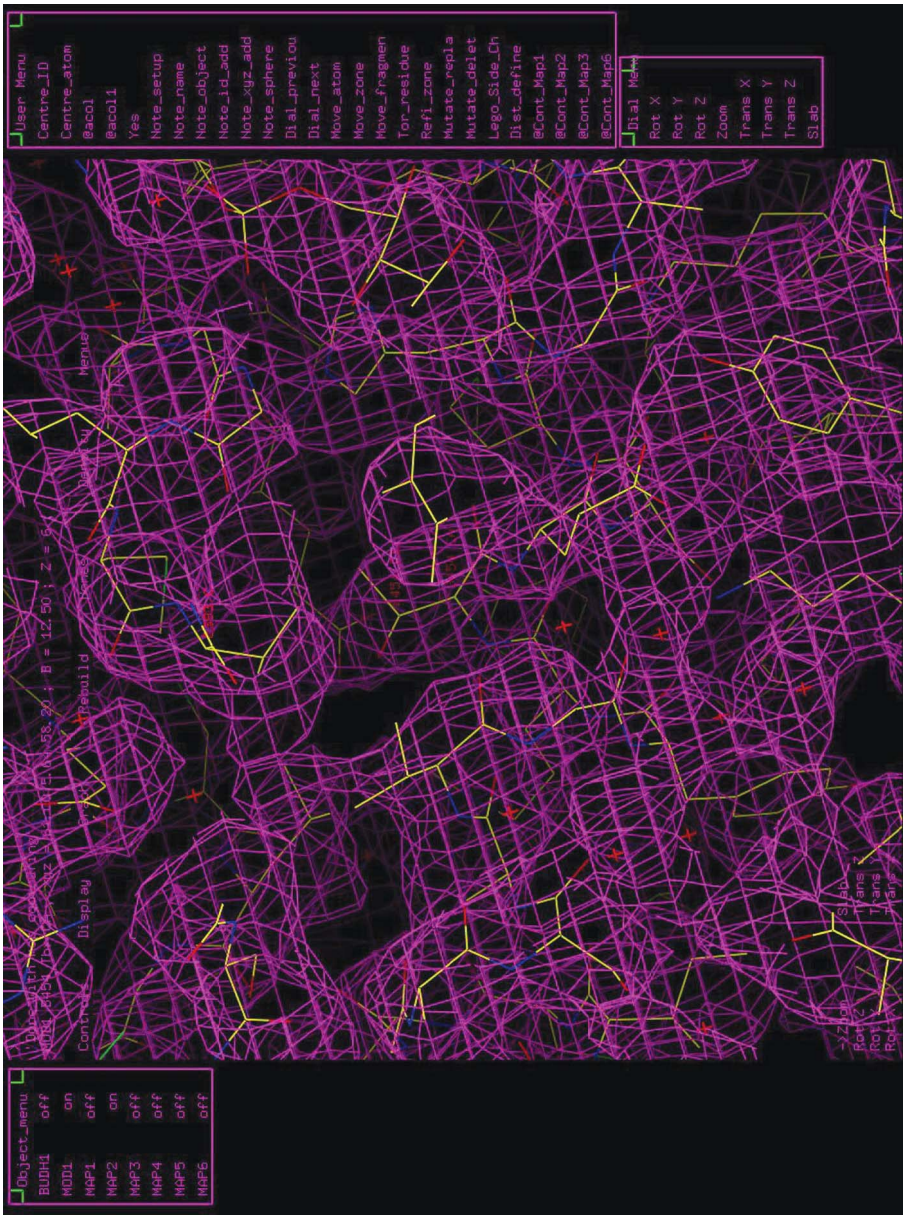


Fig. 15.1 The graphical window of program "O" as used for the model building of 4-BUDH.

data base at the start of the procedure, the reply to be made is with a "return" and further "returns" for loading several data files into the data base.

The "O" graphical window (Fig. 15.1) is opened in addition to the dialog window, where one can issue "O" commands. Initially, this window is blank (black background, color is default) with the exception of six pull-down menus. At the start, the most important bars are "Controls" and "Menus". Clicking (with the left mouse button if not specified differently) on "Menus" shows five new items. Clicking on one of the three upper items opens a small box, and clicking on the left upper corner opens a green box in the right upper corner. Clicking this box and keeping the mouse button pressed allows this box to be moved to a desired position on the graphical window. We activate the boxes for "Objects", "User Menu" and "Fake dials", and arrange them on the graphical window as depicted in Figure 15.1. Objects can be, for example, atomic models or graphical representations of electron densities. The "Dial Menu" can be used to perform rotations, translations, zooming and slabbing of the displayed objects. Initially, the box for "User Menu" is empty, but one can bring or remove individual "O" commands or text into this box with the "O" command: Menu <major_menu_name or text> <on/off> <colour>.

"O" allows read in "O" commands from a file called macros by using @ as a suffix in front of the respective file name. Figure 15.2 displays the file "start.inp", which was used to add several "O" commands and macros to the "User Menu" box and to run the setup for "Mutate" and "Lego" commands groups.

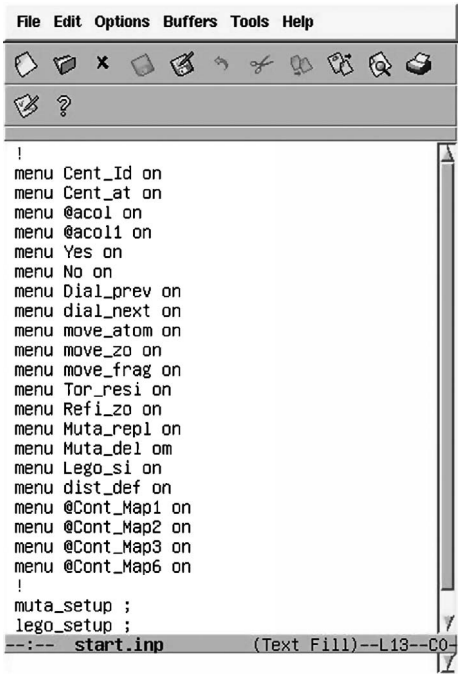


Fig. 15.2 The "Start input" file for "O" as used for model building of 4-BUDH.

Figure 15.3 shows the respective macros for @acol, @Cont_Map3 and setting symbols for @Cont_Map3.

In macro "acol", the colors for atom names starting with N, C, O, S, P and F are set. Then, object "mod1" is created for molecule "mod1" for the whole zone of the molecule. The molecule "mod1" has been previously read in with the command "pdb_read mod1.pdb mod1". The macro "Cont_Map3" is old-style "O" stuff but nevertheless very convenient in use. The files for the maps and box and contour levels (sd*) are defined via symbols given in the macro symbols3. Furthermore, the colors for the map contours are defined. The respective "symbols" macro must be read in before invoking the "Cont_Map" macro. The

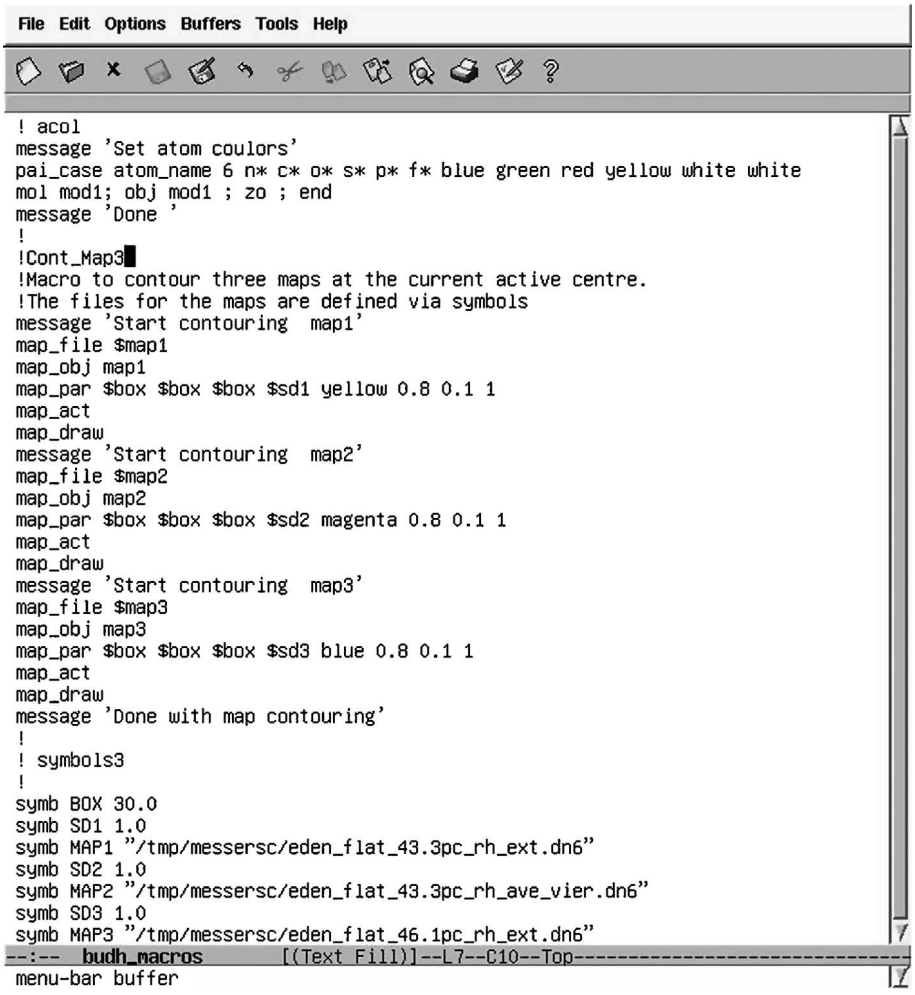


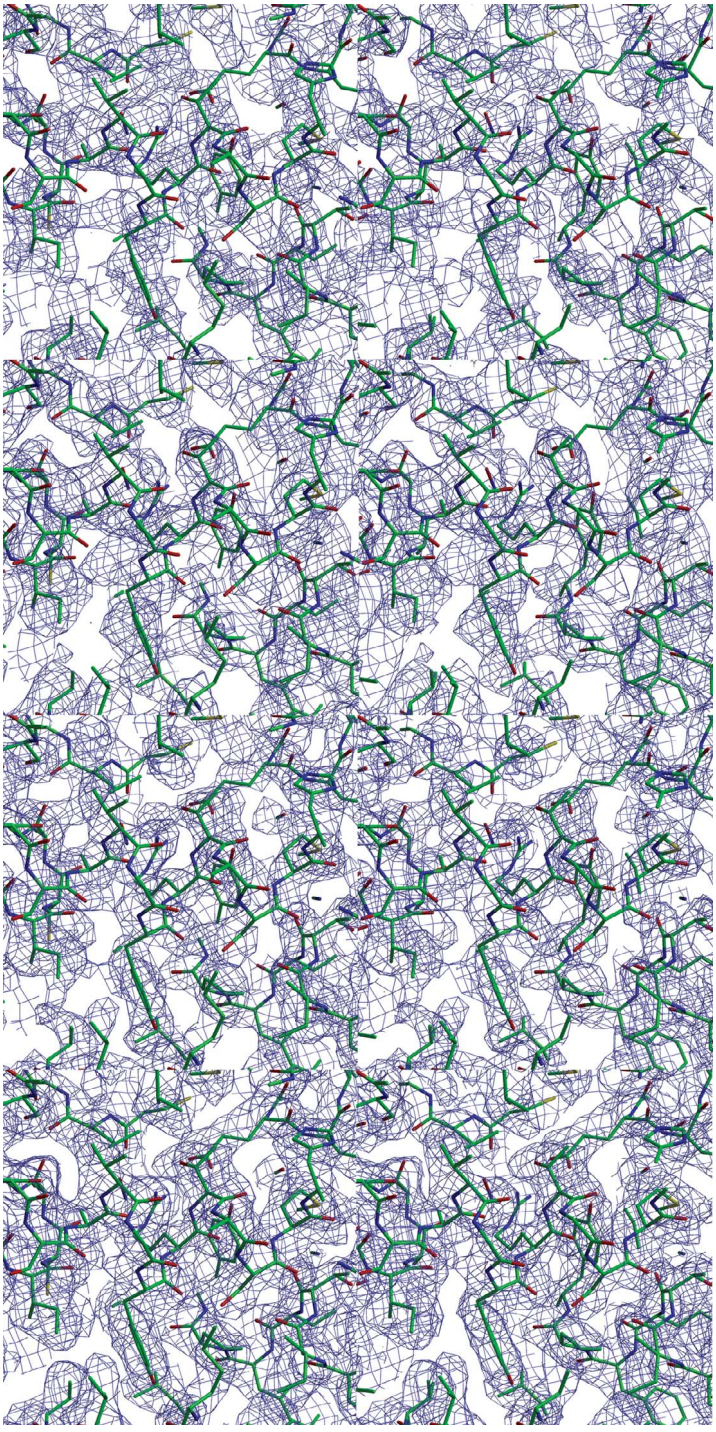
Fig. 15.3 Macros for @acol, @Cont_Map3 and setting symbols for @Cont_Map3.

new way to read in the maps is done by the "O" command "q-f <map_file> <map_name>". Five maps can be read in, and these appear in the "Density" pull down menu as Q1 to Q5. Clicking the respective item contours the relevant map, and a box is opened that allows one to adjust the radius of the contour region, the contour level, and the color of the map contours.

We can now start with the actual model building for 4-BUDH, which will be described very briefly.

At this stage, the best electron density map has been calculated from density-modified SHARP phases with four Fe-sites and subsequently averaged fourfold. A region of the map is shown in Figure 15.1, together with the final structural model. Initially, only the electron density map is displayed, of course. The map has been phased to 2.5 Å resolution, and therefore secondary structure elements such as α-helices and β-strands should be recognizable. It may be helpful to skeletonize the map and read in and display the "bones atoms" as a separate molecule. In a good-quality map the helices should show up as spirals and the strands as elongated stretches. However, large side chains such as from arginines, tyrosines, tryptophans, etc. may obstruct the appearance of the secondary elements in the skeleton. For all stages of model building with the graphics system the use of stereo glasses is strongly recommended. An α-helix is located in the central part of Figures 15.1 and 15.4. The course of this helix can be seen well, except for Figure 15.4a, where it is more difficult to recognize the feature. For the correct hand, an α-helix should be right-handed. In this orientation, the helix should go from the top of a helix-turn towards the spectator and then to the back, moving up in the back and starting a new turn. This is the case for the displayed α-helix and all other α-helices in the structure verifying the correct hand.

The next step is to put a structural model into the recognized parts of the electron density map with defined secondary structure, or where the trace of the polypeptide chain is clear. For this purpose, the USF program MOLEMAN can be used to generate coordinates for pieces of α-helices or β-strands. Program options "HELIx_generate" or "STRAnd_generate" are used, respectively. At present, we start the N-terminus of the secondary structures at (0, 0, 0). For an α-helix, one sets for the coordinates of the C-terminus (1.46*X, 0, 0) and for the β-strand (3.32*X, 0, 0), where X equals the number of residues in the structural piece. The coordinates of the generated pieces of secondary structure (polyalanine) can be written out with option "WRITe_pdb_file". Each PDB-file is read into "O" into a separate molecule and a relevant object is generated. The respective molecule is moved to the corresponding electron density area and fitted by eye to the electron density by using the "move_zone" command. The move action is terminated either by clicking "Yes" for accepting or "No" for rejecting the movement. The polarity of an α-helix or a β-strand is not easily seen at the beginning of the model building. For α-helices, the "Christmas-tree" structure of the side chains may be helpful, with large side chains pointing into the C-terminal direction of the helix. A preliminary assignment of large, medium, and small side chains and comparison of their sequence in positive and negative di-



(a)

(b)

(c)

(d)

rection with this of the amino acid sequence is useful in any case. This can be done within "O" by the old "Slider" commands, which are now part of the "Decor" commands. If the pieces of secondary structure have been correctly put into the electron density map, they can be unified into a single molecule. First, a molecule must be created that holds all amino acid residues of the 4-BUDH monomer initially as alanines. We need this single molecule at the beginning only because we build our model into a NCS-averaged map. There are several techniques to make this, for example with "sam_init_db", which needs the sequence (in our case 490 alanine residues) to be entered (in three-letter code) as a formatted O datablock.

We take one of the PDB-files of the generated secondary pieces and edit this file in the following way. We delete all residues except for the first, and rename this residue to residue name "A0" or "0" as one likes. We set all the coordinates of this residue to 1500; this assures that the residue is not displayed in "O". Next, this small PDB-file is read into "O" to a molecule with the name "BUDH", for example. Now we can use "O" command "muta_insert" to generate the whole polyalanine model. The first command would be: "muta_insert BUDH A0 A1 ala;" the second one: "muta_insert BUDH A1 A2 ala;", and so on until A490. Do not forget to save your activities by "Save_db" from time to time because something unforeseen may happen with your computer system. Now, you can move the coordinates of your fitted secondary stretches to the molecule BUDH at the correct positions. This is done by using the "merge_atoms" command. "Merge_atoms helix1 Z1 Z21 BUDH A35;" would move the coordinates from residues Z1 to Z21 of molecule helix1 to molecule BUDH at a position starting at residue A35. Now, we have the whole molecule built so far in one molecule. Be aware that the stretch of amino acids to be copied is identical to the target stretch of amino acids. Here, both sequences were polyalanines. We can try to complete the model if the electron density is good enough. Extended secondary structure and loops can be built from stretches of β-strands generated with MOLEMAN and treated in the same way as the former structural pieces. If all runs smoothly, the whole polyalanine model has been built and the side chains can be added. This is done by the commands "muta_replace" followed by "lego_side_chain". "Muta_replace BUDH A5 TYR;" would mutate residue A5 of molecule BUDH to a tyrosine. The mutated residue is displayed, but its side chain has to be fitted to the electron density. This can be assisted by "lego_side_chain". After invoking the command, one clicks on the mutated resi-

Fig. 15.4 Stereo representation of a representative section of the electron density map at different stages of phasing plus final atomic model. (a) SHARP-phases based on four Fe-sites per asymmetric unit and subsequently solvent-flattened with program DM. (b) Map (a) NCS-averaged. (c) SHARP-phases based on 16 Fe- and 16 S-sites per asymmetric unit and subsequently solvent-flattened with program DM. (d) Map (c) NCS-averaged. All maps have been contoured at 1.0 σ. The figures were produced with BOBSCRIPT (Esnouf, 1997) and RASTER3D (Merrit and Murphy, 1994).

due and several rotamers for the side chain are proposed by operating the relevant dial. The action is finished by clicking "Yes" for the best rotamer or exiting with "No" for no action. The side chain can also be fitted to the electron density with "O" command "tor_residue". The molecule will consist of stretches, which may be continuous in the amino acid sequence but have not been connected with their respective peptide bonds. This is done with the "refine" commands. First, the commands "refi_init <molecule_name>" and then "refi_gener <molecule_name> <start & end residues>" must be run. The second one must be activated after each "muta_replace" action. Sometimes one or several residues are missing in the model and these should be built in directly. For this, the following procedure is quite useful. Look for an alanine residue in the environment of the electron density map where the new residue should go to. Activate this residue with "move_zone", and then move it into the desired position in the electron density, but do not finalize the "move-zone" command at this moment. Merge the coordinates of the moved residue to the target residue in the target molecule with "merge_atoms BUDH AX AX BUDH AY". AX is the residue name of the moved residue, and AY the name of the target residue. Now, finalize the "move_zone" command with "No". The moved residue jumps back to its original position and the new residue is at its desired position. Now, redraw the molecule BUDH and make a "refi_zone" over the corresponding residue zone including several more residues on both sides of the zone.

The "lego" commands are very useful to build loop regions or extended secondary structure, but one can also construct the whole structural model. A prerequisite for this approach is the existence of a C_α-trace. The respective structure is built with help from a "database". Some of the commands are available from the pull-down menu system (Rebuild/Database). "O" can access both a main-chain database (Jones et al., 1991; Jones and Thirup, 1986) and side-chain rotamer database (Kleywegt and Jones, 1998).

The main-chain database consists of 32 protein structures that have been refined at high resolution. The nucleic acid database consists of 11 structures. The program uses the main-chain database that is appropriate for the structure. The side-chain rotamers are based on an analysis of high-resolution structures.

The rotamer database is encoded as entries in "O"'s stereochemical library and must be loaded before the relevant commands can be used. An alternative set of rotamers has been made available (The "Penultimate Rotamer Library" from Lovell et al., 2000) and is available at http://kinemage.biochem.duke.edu/databases/rotamer.html. This database has a more extensive number of conformations for each amino acid, in particular arginine and lysine residues. Special care should be taken for putative cis-peptide bonds mainly at cis-prolines, but non-proline cis-peptide bonds are also possible (Jabs et al., 1999).

The current or final model can be written out to a PDB-file and subjected to crystallographic refinement calculations. If the quality of the electron density is not so good it will be necessary to use the phase information of the model built so far for further model-building cycles. The partial model will be refined and a new electron density is calculated with these phases, or a combination of these

phases with the experimental phases. All possible density modifications should be applied to the new map. Usually, one calculates a F_{obs}–F_{calc} map apart from the $2F_{obs}$–F_{calc} map to detect gross positive and negative errors in the electron density function. It is advisable to check the quality of the geometry of the current model by producing a Ramachandran plot and a side-chain conformation analysis, which can be made with program PROCHECK (Laskowski et al., 1993). Wrong peptide orientations can be corrected with the "O" command "flip_peptide".

15.2 Introduction of the Four Fe-Sites per Fe–S-Cluster and New SHARP-Phasing for 4-BUDH

So far, we have generated two electron density maps. The first map was calculated from SHARP-phases based on four Fe-sites per asymmetric unit and subsequently solvent-flattened with the program DM. A representative section of the map showing an α-helix in the central part is displayed in Figure 15.4a. The same section for the respective NCS-averaged map is depicted in Figure 15.4b. In Figure 15.4a one can see the main parts of the electron density of the α-helix, but the electron density is interrupted several times. The NCS-averaging has strongly improved the quality of the electron density map, as can be seen from Figure 15.4(b). So far, the MAD-phasing was not optimal because each [4Fe–4S]-cluster was represented only by one Fe-site. It is, however, necessary to introduce the actual four Fe-sites per cluster.

The NCS-averaged map from Figure 15.4b has been contoured at 5.0 σ and shows electron density around the Fe–S-cluster only at this contouring level (Fig. 15.5). The coordinates of a [4Fe-4S]-cluster from the B-subunit of pyrogallol-phloroglucinol transhydroxylase (Messerschmidt et al., 2004; PDB-code 1TI2) have been read into the "O" system and fitted into this electron density manually with "move-zone". The shape of this electron density is not so well resolved to fit in the cluster exactly, but to a good approximation as the subsequent SHARP calculations showed. The actual coordinates were written out. Now, the coordinates for the Fe–S-clusters in the other NCS-related subunits had to be generated, and this was done with the USF program LSQMAN using the input file given in Figure 15.6.

The RT-operators from the NCS averaging were used, which were stored in the files "budh_1_2_imp.o", "budh_1_3_imp.o" and "budh_1_4_imp.o". The generated coordinates went into files "FeS_mol2.pdb", "FeS_mol3.pdb" and "FeS_mol4.pdb". The coordinates were transformed to fractional coordinates in the USF program MOLEMAN and a new SHARP run was performed with these 32 sites (16 Fe- and 16 S-sites). All quality factors for the MAD-phasing with these 32 sites improved, as can be seen from Table 12.1. This is also valid for the electron density map (Fig. 15.4c), the quality of which is comparable to that of the averaged map in Figure 15.4b. The new map can further be improved by NCS-averaging. In order

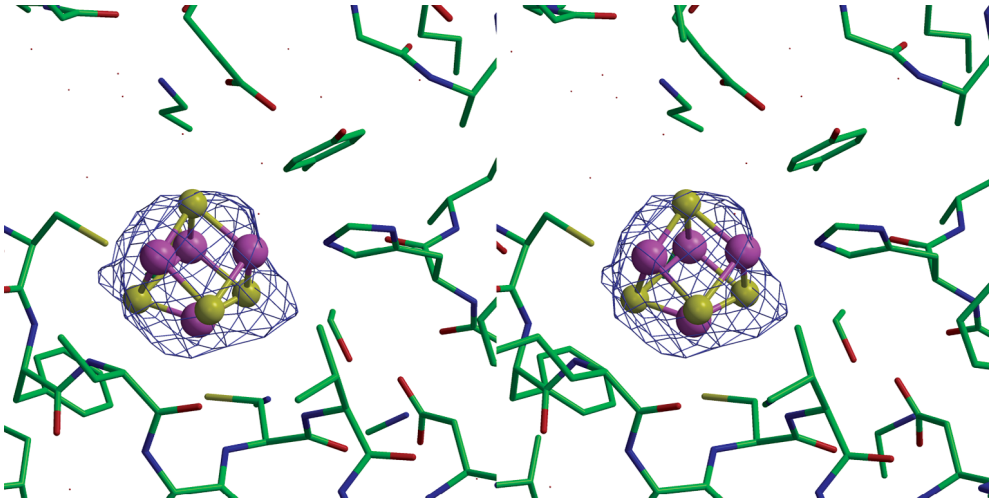


Fig. 15.5 Stereo representation of the electron density map of Figure 15.4 b around the [4Fe–4S]-cluster plus final atomic model. The map has been contoured at 5.0 σ. The figures were produced with BOBSCRIPT (Esnouf, 1997) and RASTER3D (Merrit and Murphy, 1994).

to perform an optimal averaging, the old RT-operators had to be refined with IMPROVE. The correlation factors increased as follows: monomer 1 to monomer 2, from 0.335 to 0.487; monomer 1 to monomer 3, from 0.324 to 0.480; and monomer 1 to monomer 4, from 0.356 to 0.519. The NCS-averaged map is displayed in Figure 15.4 d, and exhibits clearly the best quality. The whole model could easily be built into this electron density map.

Here, we have described a demanding case of phase calculation, density modification and model building. To an increasing degree, the quality of the initial electron density maps in many structure determinations is so good that automated model building can be applied. This can be done, as mentioned earlier in Section 7.1 with programs ARP/wARP (Lamzin et al., 2001) and SOLVE/RESOLVE (Terwilliger, 2002). The respective web pages are http://www.embl-hamburg.de/ARP/ and http://solve.lanl.gov for those who are lucky to have generated such a good electron density map.

15.3 Crystallographic Refinement and Final Steps

The next step is the crystallographic refinement of the structural model. We will not present a tutorial of this step here because there exist very good manuals and tutorials for the commonly used crystallographic programs; these include REFMAC (Murshudov et al., 1997; http://www.ysbk.york.ac.uk/~garib/refmac/)

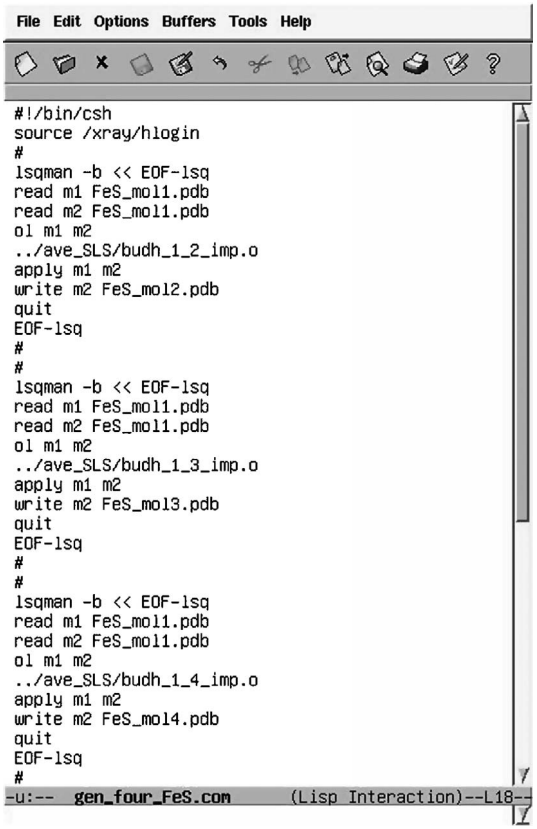


Fig. 15.6 Input file for generating the coordinates of the Fe-S-clusters in the other related subunits.

from the CCP4-suite (http://www.ccp4.ac.uk/main.html/), CNS (Brünger et al., 1998; http://cns.csb.yale.edu/v1.1/), SHELXL (Sheldrick and Schneider, 1997; http://shelx.uni-ac.gwdg.de/SHELX) and TNT (Tronrud et al., 1987; http://www.uoxray.uoregon.edu/tnt/welcome.html). We have used the program CNS for many structure refinements because it includes the simulating annealing method. Here, we will provide a short outline of using CNS for the refinement and respective input files to be referred to, as are available on the CNS web page.

In some recent cases with medium-quality diffraction data REFMAC delivered better results, however. REFMAC or SHELXL must be used in any case if the resolution is so high that a real unrestrained parameter refinement can be performed. The problem of twinned crystals has not yet been mentioned. CNS and SHELXL can, for example, refine twinned structures; this implies that the kind of twinning has been determined from the diffraction data. The detection and overcoming of twinning has been discussed by Yeates (1997), although it must

be noted that the treatment of twinned crystals is complicated and tedious, and in most cases it is better to try to grow untwinned crystals.

The crystallographic refinement procedure with CNS includes the following steps:

1. Preparation of a reflection file in CNS-format; this can be done from a CCP4 MTZ-file with CCP4 routine MTZ2VARIOUS, output option "CNS".
2. Set up a test array for cross-validation (free R) using a random selection of data. File: make_cv.inp. The percentage of the data for the test set must be given, usually values between 5% and 10%. The output reflection file has reflections for the test and working set differently flagged.
3. Generate a structure file for protein, DNA/RNA, water, ligands and/or carbohydrate. File: generate.inp. Generate needs as input apart of the PDB file of the structural model topology and parameter files for all chemical groups in the structure. The topology file contains information about the atom types, bonds, angles, dihedral angles, improper values for chirality or planarity and donors and acceptors within a special chemical group, which can be an amino acid, nucleic acid, sugar or solvent molecule, metal group or ligand. Furthermore, if groups are bonded to amino acid residues or nucleic acids, then so-called "patch residues" must be defined and be part of the respective topology file. The corresponding parameter file contains the relevant energy values for the bonds, angles, non-bonded interactions, etc. Topology and parameter files for all amino acids, nucleic acids and many ions, solvent molecules, carbohydrates and ligands are part of the CNS system.

   Our example structure 4-BUDH contains two non-standard groups, [4Fe–4S]-clusters and FAD. Topology and parameter files must be generated for them. For this purpose, the Hetero-compound Information Centre-Uppsala (HIC-Up) is very useful. Currently, this server holds about 6300 hetero compound entries, including the desired ones for the [4Fe–4S]-clusters (FS4) and FAD. The names of the hetero groups and of their atoms are consistent with those used in the Protein Data Bank. Each entry provides a PDB-file with coordinates of the compound, which can be used for placing the group into the respective position in the electron density map. Furthermore, it contains a PDB dictionary file, a CNS topology file, a CNS parameter file, an "O" dictionary entry (add to.bonds_angles datablock) and a TNT dictionary file.

   Supposed your hetero compound is not contained in this or another similar data base, it has to be generated from scratch. First, you need the atomic coordinates of the compound. A possible source is the Cambridge Structural Database (CSD: Allen et al., 1991; Kennard and Allen, 1993). If it is also not in this database, one must use a graphic molecular modeling program such as INSIGHT II (Accelrys Software Inc., San Diego, 2006; http://www.accelrys.com/products/insight) to generate the coordinates. The USF program XPLO2D can then be used to generate the topology and parameter files. If necessary, these files must be edited to meet the desired demands.

   Output files are the CNS structure file and the PDB file of the structural model suited for the subsequent CNS activities.

4. Combined simulated annealing, energy minimization, B-factor refinement, and map calculation. File: refine.inp. Input files are: PDB file of the structural model, structure file (optional), topology files, protein and DNA-RNA linkage files (supplied by the CNS system), parameters files, structure factor file, NCS-restraints/constraints file (where necessary as for 4-BUDH). Unit cell parameters, space group, high- and low-resolution limits and structure factor sigma cut-off values must be given. Parameters for the simulating annealing, energy minimization, B-factor refinement and map calculation can be usually taken as proposed in the refine.inp template file. Output files: <map_coeff_1>.map, <map_coeff_2>.map and PDB file of refined structural model. Usually, <map_coeff_1> will be $2F_{obs}–F_{calc}$ and <map_coeff_1>= $F_{obs}–F_{calc}$.

   The map files are in CNS format and can be converted to CCP4 or DN6 formats in the USF program MAPMAN.

   The logfile contains much information about the course of refinement and map calculation. There must not be warnings about close atomic distances caused by errors in the single molecules or by crystal packing. Such close contacts should be removed by a careful inspection of the structural model in the graphical modeling system because they negatively influence the energy minimization and simulated annealing procedures. If all runs smoothly, the most interesting values that the researcher is keen to obtain are the crystallographic R-factors for the test (R_{free}) and working sets. These should have decreased and the R_{free} not be more than about 5% over the working set R-factor. Normally, the refinement will be finalized after a first round. One will have to start a new cycle of model building with the refined structural model and the new improved electron density maps.
5. Pick water molecules in electron density map. File: water_pick.inp. Input files: structure file, parameter files, PDB file of refined structural model. Output files: water_pick structure file, water_pick PDB file. This routine searches the $F_{obs}–F_{calc}$-map for possible water molecules, and is a great help in establishing the solvent structure in case one did not make this by the aid of the graphics modeling program.

If initially a partial model is available only, it is advisable to run an energy minimization refinement only (template file: minimize.inp) and to start a new model building cycle. When some parts of the model are poorly defined in the electron density map, an annealed omit map may be helpful (template file: sa_omit_map.inp). A rigid body refinement (template file: rigid.inp) may be useful if a structure with low non-isomorphism (e.g., an enzyme : inhibitor complex structure with somewhat deviating unit cell parameters) to the parent structure is to be phased. One should start the rigid body refinement with the unit cell parameters of the parent structure, followed by rigid body runs with moderately changed unit cell parameters until those of the structure to be phased have been reached.

Satisfactory values for R_{free} and R depend on the resolution of the diffraction data. As a rule of thumb, the following figures are commonly accepted as satis-

factory: $R_{free} \leq 25\%$, $R \leq 20\%$. High-quality data with an overall well-defined molecular structure deliver lower values. In cases where some parts of the structure are flexible or disordered, it may happen that the values are higher. However, values about $R_{free} \sim 30\%$, $R \sim 25\%$ should not be exceeded.

If these criteria are fulfilled, then the actual crystal structure analysis has been completed. However, there remains much to do further on, as the structural model must be verified and the accuracy determined. This was described in detail in Section 7.3. If the model has successfully overcome the validation process, the coordinates and structure factor file may be submitted to the Protein Data Bank using the ADIT tool available on the RCSB web site (http://www.rcsb.org). The normal path is now to study the structure in detail and attempt to draw conclusions related to the function of the biomacromolecule. Finally, the publication reporting the results must be written. It will be necessary to prepare pictures visualizing the fold of the structure (ribbon plot), the atomic structure (optionally with relevant electron density) of selected regions of the molecule such as the active site or special loop structures, surface representations (Conolly surface, electrostatic potential, cavities), and many more. Programs for displaying 3D structures of biomacromolecules include MolScript (Kraulis, 1991; http://www.avatar.se/molscript/), BobScript (Esnouf, 1997; http://www.strubi.ox.ac.uk/bobscript/), Raster3D (Merrit and Murphy, 1994; http://skuld.bmsc.washington.edu/raster3d.html) and PyMol (DeLano, 2003; http://pymol.sourceforge.net/). Surface representations can be produced, for example, with GRASP (Nicolls et al., 1991; http://honiglab.cpmc.columbia.edu/grasp/ref.html) or DINO (2002; http://www.dino3d.org/). There are yet many other points to study in order to obtain the optimal information from a given 3D structure of a biomacromolecule, which may be complexed with another biomacromolecule, or with relevant small molecules such as substrates, products, inhibitors, effectors, or other functional molecules. The discussion of these issues is, however, beyond the scope of this book.

It is hoped that this textbook will be helpful both for students and researchers alike as a useful guide to become familiar with the methods of X-ray crystallography of biomacromolecules. These methods are capable of delivering fascinating insights into the atomic architecture and function of these molecules, which are the key players in all of life's processes.

References

Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M., Watson, D. G., *J. Chem. Inf. Comp. Sci.* **1991**, *31*, 187–204.

Brünger, A. T., Adams, P. D., Clore, G. M., Delano, W. L., Gros, P., Grossekunstleve, R. W., et al., *Acta Crystallogr.* **1998**, *D54*, 905–921.

DeLano, W. L., The PyMol Molecular Graphics System, DeLano Scientific LLC, San Carlos, CA, **2001**.

DINO: Visualizing Structural Biology (2002) http://www.dino3d.org.

Esnouf, R. M., *J. Mol. Graphics Model.* **1997**, *15*, 132–134.
Jabs, A., Weiss, M. S., Hilgenfeld, R., *J. Mol. Biol.* **1999**, *286*, 291–304.
Jones, T. A., *Acta Crystallogr.* **2004**, *D60*, 2115–2125.
Jones, T. A., Thirup, S., *EMBO J.* **1986**, *5*, 819–822.
Jones, T. A., Zou, J. Y., Cowan, S. W., Kjelgaard, M., *Acta Crystallogr.* **1991**, *A47*, 110–119.
Kennard, O., Allen, F. H., *Chem. Des. Atom. News* **1993**, *8*, 31–37.
Kleywegt, G. J., Jones, T. A., *Acta Crystallogr.* **1998**, *D54*, 1119–1131.
Kraulis, P. J., *J. Appl. Crystallogr.* **1991**, *24*, 946–950.
Lamzin, V. S., Perrakis, A., Wilson, K. S., The ARP/wARP suite for automated construction and refinement of protein models. In: Rossmann, M. G., Arnold, E. (Eds.), *International Tables for Crystallography*, Vol. F, pp. 720–722. Kluwer Academic Publishers, Dordrecht, **2001**.
Laskowski, R. A., MacArthur, M. W., Moss, D. S., Thornton, J. M., *J. Appl. Crystallogr.* **1993**, *26*, 283–291.
Lovell, S. C., Word, J. M., Richardson, J. S., Richardson, D. C., *Proteins Struct. Funct. Gen.* **2000**, *40*, 389–408.
Merrit, E. A., Murphy, M. E. P., *Acta Crystallogr.* **1994**, *B50*, 869–873.
Messerschmidt, A., Niessen, H., Abt, D., Einsle, O., Schink, B., Kroneck, P. M. H., *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 11571–11576.
Murshudov, G. N., Vagin, A. A., Dodson, E. J., *Acta Crystallogr.* **1997**, *D53*, 240–255.
Nicolls, A., Sharp, K., Honig, B., *Proteins Struct. Funct. Genet.* **1991**, *11*, 281–296.
Sheldrick, G. M., Schneider, T. R., *Methods Enzymol.* **1997**, 319–343.
Terwilliger, T. C., *Acta Crystallogr.* **2002**, *D59*, 34–44.
Tronrud, D. E., Ten Eyck, L. F., Matthews, B. W., *Acta Crystallogr.* **1987**, *A43*, 489–501.
Yeates, T. O., *Methods Enzymol.* **1997**, *276*, 344–358.

Subject Index