PART I
1

X-RAY CRYSTALLOGRAPHY OF BIOLOGICAL MACROMOLECULES: FUNDAMENTALS AND APPLICATIONS

ANTONIO L. LLAMAS-SAIZ AND MARK J. VAN RAAIJ

1.1 INTRODUCTION

X-ray crystallography is a powerful technique to determine the three-dimensional structure of any kind of molecule at atomic resolution, including that of biological macromolecules like proteins, nucleic acids, or any complex between them or with smaller compounds like ligands, drugs, cofactors, or inhibitors. The experimental result provided by this technique is the three-dimensional electron density map corresponding to the crystal subjected to the diffraction experiment. In this detailed and “amplified” image of the crystal, an atomic model for the molecules present can be built. The theoretical background involved in X-ray crystallography is very broad, covering different disciplines like mathematics, physics, chemistry, and even biology. The experimental setups can also be very complicated, like the beam lines at synchrotron installations, which include optical and experimental hutchs full of dedicated equipment. In this chapter, we will try to cover the main concepts to understand the basic theory behind an X-ray diffraction crystal structure determination and to outline, from a practical viewpoint, the principal steps in order to facilitate the interpretation of the structural determination process and the final results obtained.

1.2 FUNDAMENTALS OF X-RAY DIFFRACTION

1.2.1 X-Ray Radiation and Interaction with Matter

X-rays consist of photons from the electromagnetic spectrum with energies above ultraviolet light and below gamma radiation. The energy ranges from approximately 0.12 to 120 keV, corresponding to wavelengths between 100 and 0.1 Å, respectively (1 Å equals 0.1 nm). The most energetic X-rays, known as hard X-rays, are the ones used in crystallography for single crystal structure determination due to their penetrating abilities and due to their wavelengths that vary from 2 to 0.5 Å, similar to the shortest interatomic distances present in solid matter [1].

X-rays interact almost exclusively with the electrons of matter. They do this in different ways, via absorption, photoelectric, Compton, and Thompson scattering. Thompson scattering, also called as coherent or elastic scattering, is predominant in the X-ray diffraction pattern obtained from a crystal. It is a pure scattering interaction and deposits no energy in the scattering material. In the classical free electron model developed by J.J. Thompson in 1898, the charged particle interacts with the X-ray electromagnetic field and starts to oscillate. Consequently, it emits secondary radiation of the same wavelength (same energy) in all directions. The intensity distribution as a function of the scattering angle (angle between incident and scattered radiation) found using this classical model is comparable to that obtained from quantum mechanical calculations. As we can consider the electrons as the unique X-ray scatters in a crystal, diffraction should therefore reveal the distribution of electrons, or the electron density, of the atoms or molecules of that crystal.

1.2.2 Crystals and Symmetry

Why do we need crystals? Reconstructing the image of a single molecule using X-rays is still not possible, mainly for the following two reasons. The first one is that there is no easy way to focus X-ray-scattered beams by lenses.
The second reason is that a single molecule scatters X-rays very weakly. Having said that, X-ray diffraction from single molecules using X-ray lasers is under development [2]. Both limitations can be surpassed by the use of crystals. The crystalline state leads to the concentration of the scattered intensities for every irradiated molecule in small and well-defined regions of space (i.e., generates a diffraction pattern), thus increasing the local intensities many-fold and facilitating their measurement. After the “phase problem” has been solved (see Sections 1.2.7 and 1.3.4), the recombination of the diffracted beams is then performed by a crystallographer using crystallographic software. This is analogous to what happens in real time using an optical microscope.

The crystalline state is defined by the repetition of a single elemental unit (motif) by means of identical translations. In practice, the final result can be mono-dimensional crystals (fibers, as used in fiber diffraction), bi-dimensional crystals (a single layer of ordered molecules as used in electron diffraction), or three-dimensional crystals (used in single crystal structure determination). That means that the crystal is the convolution of a repeating motif with a periodic lattice. A crystal is formed by a motif that repeats in a perfectly regular pattern in three dimensions. Choosing any arbitrary point in the pattern and all the equivalent points related by translation, a three-dimensional lattice can be defined in which all lattice points have exactly the same environment. A fundamental difference between a crystal (the pattern) and its lattice is that the first is a continuous media (like its electron density) and the lattice is discontinuous. The point lattice is determined by all the points that correspond to successive repetitions of identical crystal components. A lattice may have additional symmetry operators besides its own translation operators and the symmetry operators belonging to the point group or space group of the corresponding crystal. For example, all crystallographic lattices are centrosymmetric.

A lattice plane can be defined for every set of three noncoplanar lattice points. All the equivalent planes in the lattice, parallel with the same periodic repetition, constitute the associated family of lattice planes. They are unambiguously named by the three Miller indices \((hkl)\) that correspond to the number of times that the planes intersect each of the three unit cell vectors \(\mathbf{a}, \mathbf{b}, \mathbf{c}\), respectively.

The unit cell is the parallelepiped built on the basis vectors, \(\mathbf{a}, \mathbf{b}, \mathbf{c}\), of a crystal lattice, which can be selected in many different ways. The most convenient way is to choose that volume enclosed by the set of three noncoplanar lattice vectors with the shortest possible lengths and sorted in a “right-handed” way. A primitive unit cell, containing only one lattice point, can always be defined. However, for symmetry reasons, basis vectors defining nonprimitive unit cells (i.e., face- or body-centered) are sometimes used instead, because they provide a more convenient coordinate system and set of basis vectors.

### TABLE 1.1 Crystal Systems and Bravais Lattices

<table>
<thead>
<tr>
<th>Crystal System</th>
<th>Lattice Centering</th>
<th>Lattice Symmetry</th>
<th>Conditions Imposed by Symmetry on Unit Cell</th>
<th>Geometry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triclinic</td>
<td>P</td>
<td>-1 (C1)</td>
<td>None</td>
<td>prismatic</td>
</tr>
<tr>
<td>Monoclinic</td>
<td>P</td>
<td>2/m (C2(^{\infty}))</td>
<td>(a = \gamma = 90^\circ)</td>
<td></td>
</tr>
<tr>
<td>Orthorhombic</td>
<td>P</td>
<td>mm2 (D(^{2}))</td>
<td>(a = \beta = \gamma = 90^\circ)</td>
<td></td>
</tr>
<tr>
<td>Tetragonal</td>
<td>P</td>
<td>4/mmm (P(^{4}))</td>
<td>(a = b = \gamma = 90^\circ)</td>
<td></td>
</tr>
<tr>
<td>Trigonal</td>
<td>R</td>
<td>-3m (D(^{3}))</td>
<td>(a = b = c)</td>
<td></td>
</tr>
<tr>
<td>Hexagonal</td>
<td>P(^{\ast})</td>
<td>6/mmm (P(^{6}))</td>
<td>(a = b = \gamma); (\gamma = 120^\circ)</td>
<td></td>
</tr>
<tr>
<td>Cubic</td>
<td>P</td>
<td>m-3m (O(^{\infty}))</td>
<td>(a = b = c); (\gamma = 90^\circ)</td>
<td></td>
</tr>
</tbody>
</table>

\(^{\ast}\)Hermann–Mauguin (and Schoenflies) symbols.

\(^{\dagger}\)The primitive hexagonal lattice is common to the trigonal and hexagonal crystal systems.

In three dimensions, seven kinds of lattices, or crystal systems, are possible: triclinic, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal, and cubic (Table 1.1). The combination of the seven crystal systems and the possibility of choosing nonprimitive unit cells give rise to 14 Bravais lattices.

The classification of a crystal into a crystal system is always determined by the symmetry of the lattice (the Laue class to which the crystal structure belongs, see next paragraph) and not to the relationships between the unit cell metric values. For example, a tetragonal unit cell will always have \(a = b\) and \(a = \beta = \gamma = 90^\circ\); however, the \(c\) axis could take any value, in most cases different from \(a\) and \(b\), but it could be equal just by chance and still belong to the tetragonal system instead of to the cubic system.

By definition, symmetry point groups apply to any object where at least one point remains invariant after the application of all its symmetry operations. Crystallographic point groups play their role in three-dimensional lattices (not in three-dimensional space in general), and in this particular case the rotations and rotoinversions allowed are restricted to \(1, 2, 3, 4, 6, \ldots -1, -2(-m), -3, -4, -6\), respectively. There are 32 crystallographic point groups (Table 1.2), also known as crystal classes. The Laue classes correspond to 11 centrosymmetric crystallographic point groups. On the
TABLE 1.2 The 32 Crystallographic Point Groups

<table>
<thead>
<tr>
<th>Laue Classes</th>
<th>Noncentrosymmetric Groups Having the Same Laue Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2/m</td>
<td>2, m</td>
</tr>
<tr>
<td>mmmm</td>
<td>222, 2/mmn</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3m</td>
<td>32, 3m</td>
</tr>
<tr>
<td>4/m</td>
<td>4, 4</td>
</tr>
<tr>
<td>4/mmm</td>
<td>422, 42/mmm</td>
</tr>
<tr>
<td>6/m</td>
<td>6, 6</td>
</tr>
<tr>
<td>6/mmm</td>
<td>622, 62/mmm</td>
</tr>
<tr>
<td>m/3</td>
<td>23</td>
</tr>
<tr>
<td>m/3m</td>
<td>432, 432</td>
</tr>
</tbody>
</table>

In bold the 11 enantiomorphic point groups.

other hand, the crystal classes that include inversion centers or mirror planes are not allowed for crystals of enantiomerically pure substances, like the biological macromolecules. Crystals of chiral molecules display only one of the 11 enantiomorphic point groups (Table 1.2).

The combination of the 32 crystallographic point groups with the 14 Bravais lattices gives rise to 73 symmorphic space groups. In a symmorphic space group, all generating symmetry operations leave at least one common point fixed, of course, with the exception of the lattice translations. To complete the 230 space groups possible in three-dimensional crystal patterns another kind of symmetry elements should be taken into account. They are the screw axes and glide planes, where the rotations or reflections are combined with translational displacements, respectively.

Crystallographic space groups apply to infinite periodic patterns. Therefore, according to the previous description, the symmetry elements of the space groups are translations, symmetry elements of the crystallographic point groups, screw axes and glide planes. In any case, the space group of a crystal structure determines its point group uniquely and not vice versa. For a complete description of all symmetry elements compatible with three-dimensional periodic patterns (crystals), see the International Tables for Crystallography, Volume A, in Reference 3. Space groups with mirror planes and/or inversion centers are not allowed for crystals of biological macromolecules, like proteins or nucleic acids, due to the enantiopure nature of these molecules. This means that there are only 65 space groups available for the enantiomorphic crystal structures of biological macromolecules.

The symmetry elements of the crystal space group operate inside the crystal unit cell; therefore, it is possible to define an “asymmetric unit” of the unit cell. The asymmetric unit is the independent fraction of the unit cell that generates the whole crystal structure once all the symmetry operations of the space group are applied. The structural description of this asymmetric unit plus the indication of the corresponding space group is all that is needed to represent the complete crystal structure (and is thus what is normally used by crystallographers, crystallographic programs, and what is deposited in databases such as the Protein Data Bank).

1.2.3 Diffraction by Crystals

Crystals are constituted by atoms; therefore let us first consider the X-ray scattering by the atomic electron cloud (considered spherically in shape in a first approximation). The scattering amplitude of an atom is called the atomic scattering factor, or form factor, $f$. It expresses the scattering power of one atom in relation of that from a free single electron, and it is calculated and averaged for spherical electron density distributions.

The values for $f$ are tabulated in the International Tables for Crystallography, Volume C, Table 6.1.1.1, page 555, in Reference 4, for each atom type as a function of $\sin \theta / \lambda$.

Usually its value is calculated using Equation 1.1 and the tabulated set of nine Cromer–Mann coefficients $a_i$, $b_i$, $c_i$ ($i = 1$ to 4) in a parameterization of the nondispersive part of the atomic scattering factor for each atom (see Table 6.1.1.4 in Reference 4). This expression is very convenient for calculation in crystal structure software suites. These values are real numbers if the X-ray wavelength is not close to an absorption edge of the atom. Near the absorption edges, the atomic scattering factors become complex numbers as expressed in Equation 1.2, where $f$ is the “normal” atomic scattering factor, $f'$ is the real part of the correction, and $f''$ is the imaginary one, which is always $\pi f'/2$ out of phase ahead of $f$ [5]. The anomalous dispersion (or more rigorously, resonant scattering) effect, far from being an inconvenience, is a very useful tool to solve crystal structures of macromolecules (see Friedel’s law description below and Section 1.3.4.3).

There is always an angular dependence for the scattering amplitude of an atom, it decays with increasing scattering angle for two reasons. The first reason is interference interactions between the scattered rays from different regions of the atomic electronic cloud. In the incident beam direction $\theta = 0$, all electrons scatter in phase, there is no decay for this reason, and the atomic scattering factor value is identical to the number of electrons in the atom. This type of decay is reflected in the tabulated values and represented with solid lines in Figure 1.1. The second source of decay is due to the atomic displacement effects that cause that the apparent size of an atom is larger than it will be at rest during the X-ray exposure time, dashed line in Figure 1.1 and Equation 1.3. The spreading of the atomic electronic cloud may be due to temperature-dependent atomic vibrations around the equilibrium position, dynamic disorder, or to the situation where equivalent atoms in different unit cells stray around different equilibrium positions. This is called static disorder and is temperature-independent. During a typical X-ray diffraction
X-RAY CRYSTALLOGRAPHY OF BIOLOGICAL MACROMOLECULES: FUNDAMENTALS AND APPLICATIONS

\[ f'(\sin \theta / \lambda) = \sum_{i=1}^{4} a_i \cdot e^{-B_i \sin \theta / \lambda} + c \]  

\[ f = f^0 + f^* + \epsilon \cdot f'' \]  

\[ f_B = f \cdot e^{-B} (\sin \theta / \lambda)^2 \]  

When considering the X-ray scattering by the whole crystal, periodicity imposes discontinuity in the resulting diffraction pattern. All scattering intensities are concentrated and magnified in well-defined directions in space where constructive interference of waves occur and are recorded as clear points in the X-ray detector. The conditions for constructive interference of the diffracted beams are defined by the Bragg’s law (Fig. 1.2) or the equivalent Laue equations.

\[ 2d \sin \theta = n \lambda \]  

\[ a \cdot (s - s_o) = h \lambda \]  

\[ b \cdot (s - s_o) = k \lambda \]  

\[ c \cdot (s - s_o) = l \lambda \]  

where \( a, b, \) and \( c \) are the unit cell vectors, \( h, k, \) and \( l \) are the Miller indices of the corresponding family of planes, and \( s \) and \( s_o \) are the unit vectors along the incident and reflected directions, respectively.

1.2.4 Real and Reciprocal Space

Given any crystal lattice in real space, it is always possible to construct its one-to-one related counterpart in reciprocal space, the reciprocal crystal lattice. The reciprocal lattice is a very convenient tool for constructing and analyzing the X-ray diffraction pattern. It is obtained by positioning its lattice points along the direction perpendicular to each family of real lattice planes and at a distance from the origin, \( d' \), equal to the inverse of the interplanar distance corresponding to this family, \( d' = 1/l \). According to this construction, each reciprocal lattice point is univocally associated to a family of lattice planes in real space. Therefore, the Miller indices of this family also correspond to the coordinates of one lattice point in the three-dimensional reciprocal lattice.

As it is clearly stated in Bragg’s law there is an inverse relation between the diffraction angle \( \theta \) and interplanar distances, \( d \). Reflections measured at higher diffraction angles correspond to shorter values of \( d \) and therefore contain structural information about the electronic density distribution at higher resolution. More detail can be seen in the electron density maps calculated with data measured up to higher diffraction angles.
1.2.5 Structure Factors

The structure factor represents the total scattered wave by all the electrons in the whole unit cell. The effective number of scattering electrons is called the structure factor, \( F \), because it depends on the structure, that is, the electronic density distribution of the atoms in the unit cell. Due to the regular periodicity in the crystals it also depends on the scattering direction. The structure factor can be regarded as the sum of the scattering by the atoms in the unit cell, taking into consideration their positions and the corresponding phase differences between the scattered waves.

\[
F(h,k,l) = \sum_{j=1}^{\text{atoms}} f_j \exp\left[2\pi i (h x_j + k y_j + l z_j)\right] \quad (1.4)
\]

\[
F(0,0,0) = \sum_{j=1}^{\text{atoms}} f_j = \text{the total scattered wave}
\]

\[
F(h,k,l) = 0 \quad \text{for } h\neq h', k\neq k', l\neq l'
\]

The electron density distribution is a periodic function; therefore, it can be represented in different ways, for example, with module and direction (phase) or as a complex number (real and imaginary part) as shown in the Argand diagram (Fig. 1.3). It is important not to confuse the mentioned “direction” of the structure factor vector in the complex space, which indicates the phase of the structure factor, with the “direction” of the diffracted X-ray beam in real space, which is determined by the crystal lattice geometry and the particular setup for the diffraction experiment.

\[
F(h,k,l) = |F(h,k,l)| \exp[\phi(h,k,l)]
\]

\[
|F(h,k,l)| = \sqrt{\sum_{j=1}^{\text{atoms}} f_j^2}
\]

\[
\phi(h,k,l) = \frac{\sum_{j=1}^{\text{atoms}} f_j \cos[2\pi (h x_j + k y_j + l z_j)]}{\sum_{j=1}^{\text{atoms}} f_j \sin[2\pi (h x_j + k y_j + l z_j)]}
\]

\[
\sin \phi = \frac{|F(h,k,l)|}{|F(0,0,0)|}
\]

\[
\cos \phi = \frac{1}{|F(0,0,0)|}
\]

The diffracted intensity is proportional to the square of the modulus of the structure factor, \( I \propto |F|^2 \). When the anomalous dispersion effect is negligible the atomic scattering factors, \( f \), of all atoms are real, and accordingly \( |F| = \sqrt{f^2} \), that is, the intensities of the \( hkl \) and \( -h-k-l \) reflections (Friedel’s pair) are equal and it is known as the Friedel’s rule. The rule does not hold for noncentrosymmetric crystals containing atoms showing anomalous dispersion, because of the imaginary part of the atomic scattering factors, \( f^* \). The difference between these intensities becomes larger when the X-ray wavelength used is close to an absorption edge of a particular atom in the crystal. Synchrotron radiation, which is a tunable X-ray source, may be used for this purpose. When the differences in intensity between both components of the Friedel’s pairs are clearly measured, the diffraction pattern reveals the symmetry of the actual point group of the crystal.

1.2.6 Fourier Synthesis and Transform

The electron density distribution is a periodic function; therefore, it can be described as a Fourier series.

\[
\rho(x,y,z) = \sum_{h} \sum_{k} \sum_{l} C_{hkl} e^{2\pi i (hx+kx+lx)} \quad (1.5)
\]

Analogously to the discrete expression for the structure factor (Eq. 1.4), it can be expressed as a continuous summation (integration) of the electron density distribution over the whole unit cell volume.

\[
F_{\text{All}} = \int \rho(x,y,z) e^{2\pi i (hx+kx+lx)} dV \quad (1.6)
\]

Substituting electronic density expression (Eq. 1.5) in Equation 1.6 and after some operations it is not so difficult to arrive to

\[
rho(x,y,z) = \frac{1}{V} \sum_{h} \sum_{k} \sum_{l} F_{\text{All}} e^{-2\pi i (hx+kx+lx)} \quad (1.7)
\]

where the structure factors are the coefficients of this summation in the Fourier expansion.

Each structure factor contains contributions from all atoms in the unit cell. Its value (module and phase) will be determined by the electron density distribution along the direction perpendicular to its associated diffracting family of planes. The reciprocal space lattice weighted by the corresponding structure factors is the Fourier transform of the electron density distribution of the crystal structure. Therefore, the reciprocal lattice construction is a very convenient representation of the diffraction pattern. To obtain this information, every measured diffracted intensity has to be processed (see Section 1.3.3) to get the structure factor module after

\[
\begin{align*}
|F(h,k,l)| & = \frac{1}{V} \sum_{j=1}^{\text{atoms}} f_j \exp[\phi(h,k,l)] \\
\phi(h,k,l) & = \frac{\sum_{j=1}^{\text{atoms}} f_j \cos[2\pi (h x_j + k y_j + l z_j)]}{\sum_{j=1}^{\text{atoms}} f_j \sin[2\pi (h x_j + k y_j + l z_j)]}
\end{align*}
\]
normalization and correction for Lorentz, polarization, and absorption effects.

1.2.7 The Phase Problem

Only X-ray-diffracted intensities can be measured, and from them only the value of the amplitude (module) can be estimated. No direct information about the phase can be recorded in the X-ray diffraction experiment, and the reciprocal space lattice could only be weighted by structure factor amplitudes. To calculate the Fourier transform and obtain the three-dimensional electron density map of the crystal, the value for the phase of each reflection is needed. The crystallographer must obtain the phase angles from further experimentation as described in Section 1.3.4. This is what is called the "phase problem" in crystal structure determination [6, 7].

1.3 THE STRUCTURE DETERMINATION PROCESS

Determining the structure of a macromolecule is a process that consists of various steps, comprising many different techniques. The macromolecule, or complex of macromolecules, may have to be expressed in a suitable system if it cannot be isolated from natural sources. For this, a suitable expression vector will need to be constructed, involving genetic engineering and/or cloning. The molecule or complex of interest will have to be isolated and purified, either from its natural source or from the expression host in sufficient amounts, usually several to many milligrams. Then, many different crystallization trials are performed. When crystals are obtained, they have to be manipulated to allow data collection, and where necessary, heavy atom derivatives may need to be prepared. Cocrystralization or crystal soaking experiments with natural or artificial ligands may also be performed. This part takes place in the laboratory, that is, in vitro. Data processing, structure determination, model construction, refinement, and validation take place in silico, using specialized computer programs developed to such end. All these steps are discussed below.

1.3.1 Sample Production and Conditioning

High-quality samples may be obtained by careful purification from natural sources in which the macromolecule or complex of interest is present in sufficient amounts and at high enough concentration to make purification feasible and worthwhile. Examples are myoglobin from sperm whale meat [8], hemoglobin from blood [9], elongation factor Tu and ribosomes from bacteria [10, 11], F$_1$-ATPase from beef hearts [12], and light harvesting center from spinach leaves [13]. However, in many other cases, the macromolecule or complex of interest needs to be expressed in bacteria, yeast, insect cells, or mammalian cells.

1.3.1.1 Protein Expression in Bacteria

For expression in prokaryotic systems (most often the bacterium Escherichia coli), expression vectors have to be constructed. Usually, expression plasmids are used. Plasmids are small circular DNAs that replicate in the bacterium independently from the chromosome. To select for bacteria containing the plasmid during cultivation, plasmids contain a gene encoding an antibiotic that confers resistance to a certain antibiotic. For example, they may encode a gene for beta-lactamase, which hydrolyses ampicillin and carbenicillin. Other commonly used antibiotics are kanamycin, streptomycin, and chloramphenicol with their corresponding resistance-conferring genes. Positive selection of plasmid-containing bacteria is necessary because, without selection, bacteria without incorporated plasmid will inevitably have a growth advantage due to less energy expenditure and they will thus outgrow plasmid-containing ones.

To allow for replication in bacteria, plasmids must contain an origin of replication, the type of which also determines whether the plasmid is present at higher or lower copy numbers. In many cases, high plasmid copy numbers are desirable, because it facilitates plasmid purification and allows for the expression of high amounts of protein in less time. However, in cases where the protein folding rate is limiting, it may be preferable to have lower plasmid copy numbers, leading to less rapid protein expression and thus giving more time to the expressed proteins to fold correctly. Growing the cultures at lower temperatures may also promote correct folding.

The promoter and its location upstream of the gene to be expressed included in the expression vector determine the amount of messenger RNA that will be produced and thus, indirectly, the rate and amount of protein that will be expressed. In principle, constitutive expression may be employed, but unless the expressed protein is useful for the expression host (e.g., a chaperone), the extra expenditure of energy to produce the protein will be disadvantageous and mutants that do not express the protein will accumulate during repeated growth/dilution cycles. Therefore, several inducible expression systems have been developed. Many use the P$_{lac}$, P$_{lac}$, or P$_{lac}$ promoters, inducible with the lactose analogue isopropyl-beta-thiogalactoside [14, 15]. Another popular system uses the P$_{UV5}$ promoter, the late promoter of bacteriophage T7 [16]. In this case, first T7 RNA polymerase has to be produced, which is usually achieved using an expression host that contains a lambda lysogen called DE3, which encodes T7 RNA polymerase under the control of the isopropyl-beta-thiogalactoside-inducible lam-UV5 promoter. The T7 RNA polymerase then produces the messenger RNA of interest.

Most inducible systems allow some protein expression even before induction. This means that if the protein or complex to be expressed is toxic to the host cells, a system with strong repression before induction must be used. An example of such a system uses the P$_{BAD}$ promoter of the E. coli
arabinose operon and its regulatory gene araC, allowing strong repression in the absence of l-arabinose (and even stronger repression if glucose is added to the culture media) and high levels of messenger RNA generation after induction with l-arabinose [17]. In case the protein to be expressed contains cysteine bonds, expression in the reducing bacterial cytoplasm may lead to incorrectly folded protein. In this case, the protein to be expressed may be directed to the less-reducing bacterial periplasm compartment via an N-terminal signal peptide or bacterial strains mutated in thioredoxin reductase (trxB) and/or glutathione reductase (gor) may be used (like the E. coli Origami strain). For some proteins, coexpression with a specific chaperone or chaperonins may be necessary for correct folding [18]. They may be encoded on the same plasmid or another plasmid to be cotransformed into the bacteria or their coding sequence may be integrated into the host genome. Another reason for low expression levels may be that the heterologous gene contains a codon that is very rare in the bacterium used. Use of a strain overexpressing rare tRNA species may resolve this problem (for instance the E. coli Rosetta strain).

When the object of interest is a protein complex, proteins may be mixed after purification or after expression, and the resulting complex is purified directly. Proteins may also be coexpressed using expression vectors encoding two or more proteins or by the use of multiple expression vectors in the same bacterial host. These multiple expression vectors should be compatible and encode different antibiotic resistance genes, so that selection using the relevant antibiotics simultaneously forces the bacteria to maintain all the plasmids. Terpe [19] has written a short but comprehensive review of commonly used bacterial expression systems.

1.3.1.2 Protein Expression in Eukaryotic Systems

Not all eukaryotic proteins fold correctly in prokaryotic expression systems, in which case expression in eukaryotic systems may be tried. Eukaryotic systems may also be necessary if the expressed protein is to contain certain posttranslational modifications. As a single-celled and innocuous organism, the yeast Saccharomyces cerevisiae has been most extensively studied for protein production (reviewed in Reference 20). Expression plasmids have been developed with sequences for propagation in E. coli for DNA amplification and in yeast for protein expression experiments, including yeast promoters and terminators for the production of messenger RNA. Chromosomal integration of a suitable protein expression cassette is also an option, as plasmids are not always stably maintained in yeast cells. Another yeast species, Pichia pastoris, is noted for its high endogenous protein production capacity and is also used routinely [21]. In P. pastoris, expression vectors that integrate into the genome appear to be the norm. In both yeast systems, the proteins to be expressed may be directed to the medium or allowed to accumulate intracellularly.

Cloning the gene to be expressed in a viral vector and infecting eukaryotic cells with the resulting viruses is also a system that can produce high yields of protein. The system that is developed most for protein expression is infecting insect (lepidopteran) cells using recombinant baculovirus [22]. Recombinant baculoviruses are constructed by replacing the polyhedrin gene by a gene encoding the protein of interest. Expression is controlled by the strong late polh promoter, which thus allows the production of the recombinant protein at high yield. In vitro, polyhedrin is produced at high amounts (up to 50% of the total infected larva protein mass) and is necessary to form occluded virus, which can survive in the environment until uptake by a new feeding caterpillar. In vitro, polyhedrin is not necessary for virus survival because budded virus can readily infect cultured insect cells and replicate in them. Methods to express multiple proteins to form protein complexes in the baculovirus/insect system have been developed [23]. Other viral systems that have been developed for protein expression include vaccinia virus [24], which allows transient expression in human cell lines (such as HeLa cells). Usually, the P7 promoter is used, and the T7 RNA polymerase necessary for this is either constitutively expressed in the cell line used or included in the recombinant vaccinia virus vector.

The DNA containing the gene for the protein to be expressed can also be transferred into eukaryotic cells by transfection. For this, a suitable DNA vector is usually constructed as a plasmid in E. coli and transfected into mammalian cells by electroporation or using cationic lipids (lipofection) for transient expression [25]. Popular cell lines are HEK293 [26], derived from human embryonic kidney and CHO, derived from Chinese hamster ovary. Cells that have incorporated the DNA into their genome and express the recombinant protein in a stable manner may be selected.

1.3.1.3 Cell-Free Protein Expression

In case the protein to be expressed is toxic for living cells or very prone to degradation, a cell-free in vitro translation system may be a viable, albeit more expensive, solution. For in vitro protein expression, first messenger RNA must be produced by in vitro transcription. Bacteriophage T7 RNA polymerase may be used for this. In this system, the gene of interest is cloned behind a T7 promoter, allowing large amounts of messenger RNA to be produced when DNA, nucleoside triphosphates, and T7 RNA polymerase are mixed. For the translation step, apart from the messenger RNA, many other components are necessary (initiation factors, ribosomes, transfer RNAs, elongation factors, amino acids, ATP and GTP, termination factors, ions), so that usually cell extracts are used that contain all of them. Examples are rabbit reticulocyte lysate and wheat germ extract. Coupled systems are available in which the transcription and translation steps occur in the same tube, either by the same cell extract (such as an E. coli extract) or by mixing the components necessary for the two steps. An
advantage of in vitro systems is that ligands or other protein interaction partners may be added, which in vivo may not be taken up by the cell or be degraded by living cells before they can interact with the expressed protein. These interaction partners may make the protein more soluble and/or more stable [27]. More details about cell-free expression systems are available in specialist books, such as that by Spirin and Schwitz [28].

1.3.1.4 Production of Nucleic Acids For the study of DNA and RNA structure (alone and in complex with nucleic-acid-binding proteins like transcription factors and restriction enzymes), crystalization-quality nucleic acids will need to be obtained. DNA molecules may be synthesized chemically, and many companies provide oligonucleotide synthesis services. RNA oligos may also be synthesized but are more costly and difficult to produce due to the necessity of protecting the extra 2'-hydroxyl group. RNAs may also be produced in the lab by in vitro transcription [29]. The template may be a pair of complementary DNA oligonucleotides encoding the T7 promoter and the sequence of the RNA to be produced downstream from it. A gene encoding the RNA molecule to be produced may also be cloned into a plasmid under control of the T7 promoter, the plasmid amplified in E. coli and purified in large amounts. After linearization of the plasmid with an efficient restriction enzyme, T7 RNA polymerase is added along with nucleoside triphosphates, leading to the production of large amounts of RNA. For efficient transcription by T7 RNA polymerase, the first few bases of the RNA to be produced should be purines, while the sequence of the 3' end is determined by the restriction enzyme used. To avoid these restrictions, a 5' cis-acting autoleaving hammerhead ribozyme may be encoded 5' and 3' of the sequence to be produced [30]. These authors also pioneered the use of the restriction enzyme BamAI that cleaves 5' to its recognition site to digest the template DNA prior to transcription. In this way, no restrictions exist for the sequence at the 3' end of the desired RNA.

1.3.1.5 Purification and Conditioning After production, the macromolecules or complexes to be crystallized need to be purified. Oligonucleotides, where irreversible unfolding is less of a problem than for proteins, may be purified by polyacrylamide gel electrophoresis or high-performance liquid chromatography. If the protein is present in the cultivation medium, it may be purified directly from it after removal of the expression host cells by centrifugation. This has the advantage of a relative absence of insoluble contaminants but the disadvantage of a relatively large volume. If the protein is produced intracellularly or is to be purified from a natural tissue source (e.g., meat or spinach leaves, see Section 1.3.1), a crude extract will need to be prepared. Cells will need to be broken by grinding, sonication, treatment with a hypotonic solution, detergent treatment, or treatment with a cell-wall destroying enzyme like lysozyme. In the case of soluble proteins, cell debris may be removed by centrifugation and the protein purified from the soluble fraction, while in the case of membrane proteins, the protein may be extracted from the membrane with detergents. If the protein of interest is expressed as inclusion bodies, these may be purified by differential centrifugation and sucrose gradient centrifugation and the protein refolded from these inclusion bodies [31]. However, protein refolding is often not straightforward and there is no guarantee of success.

To facilitate purification, proteins may be expressed with purification tags or as fusion proteins. The first purification step may then be performed using affinity chromatography, examples are metal affinity chromatography for proteins containing an oligohistidine sequence, a matrix with a modified streptavidin for proteins with a streptavidin-recognizing octapeptide, amylose-resin chromatography for proteins expressed as maltose-binding protein fusions, or glutathione affinity chromatography for protein containing a glutathione-S-transferase tag. Maltose-binding protein and glutathione-S-transferase have the additional advantage that they may help the target protein stay soluble during expression, although for crystallization such a large fusion partner is likely to be detrimental and would have to be removed (usually by including a specific protease site between the two fusion partners). If no purification tag is present, usually some bulk fractionation step needs to be performed before proceeding to more traditional column chromatography steps. These may include ammonium sulfate precipitation, streptomycin sulfate precipitation to remove nucleic acids, or sucrose gradient centrifugations to isolate large complexes. Then, purification takes place using anion and/or cation exchange chromatography and size exclusion chromatography (often as a final “polishing” step). It should be stressed that no universal purification protocols are available and specialized schemes have to be developed for each particular protein.

During and after purification, the identity and state of the sample should be verified. In the case of proteins, N-terminal sequence analysis (Edman degradation) and mass spectrometry can be used to verify the identity of the protein and to verify that the N-terminus (and sometimes C-terminus) are as expected. In the case of enzymes and macromolecules that bind specific ligands, activity and binding assays may be performed to verify identity and correct folding.

For successful crystallization, it is usually necessary to concentrate the purified macromolecule to values of more than 10 mg/mL. Although proteins have been successfully crystallized from samples at 2 mg/mL or less, a higher concentration increases the chances of success, and if the protein is maintained soluble at 20, 50, or even 100 mg/mL, crystallization trials may be setup at these higher concentrations. Concentration of macromolecular samples may be achieved by filtration using membranes through which the protein does
not pass. The necessary pressure to force the buffer through the membrane may be provided by centrifugation or pressurized nitrogen or air. Alternative methods include protein precipitation by ammonium sulfate followed by dialysis or by covering a dialysis tube containing the sample with polyethylene glycol powder, removing solvent from the sample but retaining the macromolecule in the tube, optionally followed by dialysis.

Crystals consist of regularly repeating units of the same molecule or complex, each in the same conformation. In order for a sample to successfully crystallize, purity is very important. Therefore the minimum amount of buffer components to keep the protein stable should be included—in fact, many macromolecules are stable in water alone, and the purification buffer can be exchanged for water or the minimum buffer in the last concentration or dialysis step. The chemical purity of the macromolecule or complex may be assessed using denaturing gel electrophoresis. This should also reveal if the protein is intact or whether proteolysis may have occurred during expression and purification.

While chemical purity is necessary, it is not sufficient; conformational homogeneity is just as important. Typical causes of conformational heterogeneity may be partial and unspecific aggregation, unfolding or flexible domains. The aggregation state of the protein may be investigated by native gel electrophoresis, size exclusion chromatography, dynamic light scattering, or analytical ultracentrifugation. The fact that a protein forms oligomers is not necessarily a problem, as long as it forms a homogeneous population of them, leading to a monodispersive sample. Certain proteins may need to form specific oligomers to perform their natural function and may not even be as stable as monomers. If the macromolecule or complex is large enough, it may be useful to observe single particles by electron microscopy, which may quickly reveal large differences in conformation or oligomerization state using only small amounts of sample. Native gel electrophoresis or isoelectric focusing may also reveal multiple charge states for the macromolecule. If this happens, these may need to be separated by ion exchange chromatography or preparative isoelectric focusing.

To have a reasonable chance of crystallizing, the macromolecule or complex of interest should be folded correctly. While many unfolded proteins aggregate unspecifically and often even precipitate, some proteins may be perfectly soluble and monomeric, even when unfolded. The folding degree of a protein may be judged by NMR spectroscopy, a folded protein should have a more dispersed set of amide protein resonances when compared to unfolded, random coil, proteins (see also Chapter 2). If it is suspected that the macromolecule has disordered loops or larger flexible domains, it may be necessary to remove these by limited proteolysis or by redesigning the expression vector. A specific ligand or inhibitor may also be included to try and lock the protein, the nucleic acid, or complex into a unique conformation.

1.3.2 Crystalization

Several different methods exist for obtaining crystals of macromolecules. In most of them, the solution containing the macromolecules (the mother liquor) is mixed with a similar volume of precipitation solution and allowed to equilibrate with a larger volume of the same precipitant solution. Equilibration by vapor diffusion is the most commonly used method. Traditionally, this was (and is) performed by the hanging drop method, placing the drop of mother liquor on a siliconized microscope cover slip and inverting this cover slip over a well with precipitant solution in a Linbro plate. The borders of the well are sealed with mineral oil or vacuum grease. Currently, sitting drop vapor diffusion experiments are becoming more popular because of their relative ease of setup, ease of crystal harvesting, and suitability for automatization. Sitting drop vapor diffusion experiments can be sealed with extraclear tape, which permits opening individual wells by carefully removing the tape only from that well and rescaling with a piece of the same tape.

Some proteins are sensitive to air, and although vapor diffusion experiments can be setup under a nitrogen atmosphere to prevent oxidation, dialysis may be a better option [12]. Microdialysis buttons are available for small volumes (5–350 μL of mother liquor), although these are still an order of magnitude larger than the volumes used in vapor diffusion or microbatch experiments (see next paragraph). The buttons are covered with a piece of dialysis membrane kept in place with a rubber o-ring and incubated in a vial with a large volume of precipitant solution. A further advantage of this method is that after crystal growth, ligands, cryoprotectant, and other components can be introduced into the mother liquor without disturbing the crystals by adding them to the precipitant solution or exchanging the precipitant solution and waiting for equilibration.

Macromolecules can also be crystallized in batch, by simply mixing a concentrated solution of them with precipitant solution and waiting. In microbatch experiments, protein solution is directly mixed with precipitant solution and allowed to equilibrate with a larger volume of precipitation solution and allowed for slow evaporation of aqueous solvent through the oil layer. A percentage of silicon oil can be mixed in with the mineral oil if faster evaporation is desired. This is often done in Terasaki plates, which contain 60 or 72 small wells.

Free interface diffusion is another commonly used technique [32]. The solution containing the concentrated macromolecules is brought into direct contact with the precipitant solution in a capillary and slow free diffusion is allowed to take place through the small contact surface. The concentration gradient that forms along the capillary allows sampling of a larger fraction of crystallization space in a smaller number of experiments.

THE STRUCTURE DETERMINATION PROCESS 11
Crystalization robots can significantly expedite the crystallization process, eliminating a lot of tedious manipulations and allowing for small-volume drops (typically 50 μL). There are robots specialized in microbatch experiments or sitting-drop vapor diffusion, but multipurpose ones are also available that can also perform hanging-drop vapor diffusion experiments. Robots generally use 96-well plates, with the possibility of multiple crystallization drops per well.

A typical initial screen consists of one or more 96-well plates with very different conditions [33, 34], and if possible, the same experiments are incubated at different temperatures (e.g., at 20°C and 5°C). Incubation should be in low-vibration conditions. If crystals are obtained, they are measured to confirm they are protein, not salt or another small-molecule additive, and to assess their diffraction limit and quality. If crystalline precipitates are obtained, further screens are performed around these conditions to see if crystals can be obtained. At the same time, it is worth carefully examining the cloning, expression, and purification strategy to see if improvements in protein purity and conformational homogeneity can be obtained (see Section 1.3.1.5). In addition to these initial more-or-less random screens, it is worth screening common precipitants such as ammonium sulfate and polyethylene glycol at different concentrations, pH, and temperatures. Precipitant solutions should be prepared using high-grade chemicals. Other parameters that may be varied to obtain crystals or improve their size and quality are initial protein concentration, drop size, and the ratio of protein solution to precipitation solution in the drop. Additives of different classes may be tried, such as multivalent cations, common salts, chaotropes, reducing agents, polyamines, and organic molecules.

The results of crystallization experiments include clear drops and precipitates due to unspecific protein aggregation. In these cases, future experiments in which the precipitant concentration is increased or decreased, respectively, may yield more promising results. Phase separation in which the protein concentrates in an organic phase may also be observed, and sometimes protein crystals nucleate on the edges of such organic phases. Crystalline precipitates may form due to excessive nucleation or inversely, clusters of crystals due to insufficient nucleation sites. Sometimes, crystals or crystal fragments useful for diffraction experiments may be separated from these clusters. Single crystals may also be observed. Often, crystal growth is not equally efficient in all three dimensions and needle- or plate-shaped crystals result, but if the conditions are just right, crystals with sizes of 10–100 μm in all three dimensions may be obtained. Where crystals are too small, seeding drops with preformed microcrystals may lead to growth of larger crystals [35, 36]. Seeding may also improve crystal qualities other than size. For more complete texts on protein crystallization, textbooks are available [37–39].

1.3.3 Data Collection and Processing

The first step of data collection is the recovery of the fragile crystals from the crystallization setup. For room temperature data collection, they may be carefully transferred to a quartz capillary and mounted in conditions in which the crystal will not dry up or be able to attract moisture from the surrounding atmosphere and dissolve. They can also be picked up with a nylon or plastic microloop about the same size as the crystal. The loop is then covered with a plastic hood filled with a drop of mother liquor. To prolong crystal life, a crystal can also be briefly incubated in a suitable cryoprotectant, and in this case, they can either be flash-frozen at 100 K inside a nitrogen gas stream or in liquid nitrogen [40]. If data collection is then performed at 90–120 K, a significant increase in crystal lifetime can be obtained as radiation damage decreases at lower temperature [41].

The most common strategy setup used nowadays to measure X-ray diffraction intensities is the oscillation method. Consecutive images are recorded for small rotation angles (0.25° to 2°) around an axis perpendicular to the incident X-ray beam [42]. Depending on the space group of the crystals obtained and the structure solution method that is to be used, somewhat different data collection procedures will need to be employed. In all cases, complete datasets are necessary, and if the diffraction data anomalous signal is to be exploited, Friedel’s pairs will have to be collected for each reflection at high multiplicity. This is because the anomalous intensity differences between Friedel’s pairs are generally small compared to the diffraction intensities. For high-symmetry space groups, a relatively small fraction of reciprocal space needs to be explored, while for lower-symmetry space groups, a larger fraction of reciprocal space will need to be covered, that is, more images per dataset will have to be collected. For structure solution by molecular replacement or isomorphous replacement methods (see Section 1.3.4), high multiplicity is not a necessity (although it is always an advantage), while for anomalous dispersion methods it is very important. High-multiplicity datasets will require longer data collection times, while at the same time radiation damage will have to be avoided [43]. Therefore, to allow successful structure solution, at times higher resolution data will have to be sacrificed (i.e., less exposure time per image) for data completeness and/or multiplicity. Once the structure is solved and more crystals are available, one can always attempt to collect a complete higher resolution dataset for the final refinement of the structure. Completeness means that as many as possible reflections for this particular crystal structure are well-measured. A common mistake is to overexpose crystals in order to achieve the highest possible resolution, leading to overloading low-resolution reflections. In some cases this problem is best overcome by merging two datasets measured at low- and high-beam intensity or exposure time.
1.3.4 Structure Determination

Crystal structure determination is basically the resolution of "the phase problem"; different methods have been developed to estimate phase values. The first protein structures, myoglobin and hemoglobin, were solved using multiple isomorphous replacement (MIR) [44,45]. These days, in many cases homologous protein structures are available and molecular replacement can be successful. The development of variable-wavelength X-ray sources at synchrotrons has led to the increased use of anomalous dispersion-based methods to solve crystal structures.

1.3.4.1 Molecular Replacement

New structures isomorphous to already known structures may be solved by Fourier synthesis using the phases calculated from the previous structure combined with the diffraction intensity data collected from the new crystal. Isomorphous means both crystals have the same space group, very similar cell parameters, and the same orientation of the molecules in the asymmetric unit. Common examples are solving the structure of the same protein with a new ligand or with a point mutation in its sequence.

If a similar structure is known from a crystal that belongs to a different space group and/or with significantly different lattice parameters, structure solution using the molecular replacement technique may be possible [46]. For molecular replacement to work, the search model will have to be a significant fraction of the total structure to be solved and sufficiently similar in structure. In general, if the protein sequence identity is 25–30% or more, a reasonable chance of success can be expected. However, it should be remembered that sequence similarity is not what is important, but structure similarity is, which means that sometimes molecular replacement can be successful with search models with less sequence identity or fail with search models with more sequence identity than the mentioned cutoff.

For structure solution by molecular replacement, in most space groups six parameters have to be determined: three rotation angles and three translation shifts to be applied to the search model. Some exceptions are triclinic space groups (only the three rotation parameters, no translation parameters) and monoclinic space groups (three rotation angles and two translation shifts). A full six-parameter search can be performed, but it is computationally very intensive, which is why most molecular replacement protocols first determine the rotation angles, then the translation parameters, and then perform a quick rigid body refinement to optimize all six parameters at once (fitting).

Patterson functions, which can be calculated without phases, are calculated for the model and for the experimental data. Self-vectors in the Patterson function (intramolecular vectors from one atom in the molecule to another atom in the same molecule) depend on the orientation of the molecule and are used in the rotation function. The three angles where the self-vector Patterson functions are most similar to each other determine the orientation of the search molecule in the cell. Cross-vectors in the Patterson function (intermolecular vectors from the atom in the molecule to the equivalent ones in the other molecules) depend on both the orientation of the molecule and on its position in the cell. So, once the orientation is known, cross-vectors can be exploited in the translation function to determine the translational shifts.

Computer programs used for molecular replacement include AMORE [47], MOLREP [48], and PHASER [49].

1.3.4.2 Direct Methods

In case very high-resolution data can be obtained and not too many atoms are present in the asymmetric unit, structure solution by direct methods may be possible. Limit estimates are around 1.2 Å or better for the resolution and up to 200–1000 nonhydrogen atoms in the asymmetric unit, which is rare for crystals of macromolecules. Direct methods are based on mathematical relationships among certain combinations of phases. Cosine values of phase combinations known as triplet structure invariants can be reliably estimated if measured intensities are large (i.e., good diffraction) and the number of atoms in the asymmetric unit is small. Multiple sets of trial phases are constructed, and each phase is refined using these mathematical relationships. In favorable cases, initial phase estimates converge toward a complete set of phases with small phase errors. Another approach is to try out random arrangements of atoms in the asymmetric unit, simulate their diffraction patterns, and compare these simulated patterns with those obtained from the crystals. Correct solutions should have high correlations between the simulated and the observed diffraction patterns. Even if only physically possible arrangements of atoms are tried, the number of trial arrangements to test quickly gets too large for big molecules. However, even for very large macromolecules or complexes, direct methods are often used to locate the limited number of heavy atoms in derivative datasets, see Section 1.3.4.3. Programs for direct methods include SNB [50] and SHELX [51].

1.3.4.3 Isomorphous Replacement and Anomalous Dispersion

If molecular replacement is not successful, heavy atom derivatives will have to be produced for structure solution by multiple isomorphous replacement (MIR), single isomorphous replacement using anomalous signal (SIRAS), multiwavelength anomalous diffraction (MAD) [52], or single-wavelength anomalous diffraction (SAD) [53]. Common derivatives are mercury compounds, which bind covalently to cysteine residues and are especially useful for MIR or SIRAS, or selenomethionine derivatives, especially useful for the MAD method [54]. Specific radiation damage may also be used to solve macromolecular structures [55]. Heavy atoms naturally present in some proteins (i.e., metal-binding proteins) may also be used for phasing, and in favorable cases,
the anomalous dispersion properties of sulfur (proteins) and phosphate (nucleic acids) may help determining phases.

Heavy atoms are generally introduced into preformed protein crystals by soaking techniques [56] although cocrystallization is also a possibility. Seleniummethionine can be introduced into proteins instead of methionine by growing methionine-auxotroph bacteria in expression cultures in the presence of selenomethionine [57] or by the inhibition of the methionine synthesis pathway and provision of the necessary amino acids and selenomethionine in expression cultures. If no cysteines or methionines are present in the natural sequence, these can be introduced by site-directed mutagenesis.

The isomorphous replacement technique uses the intensity differences between equivalent reflections of datasets measured from crystals of the “native” macromolecule and crystals of the same macromolecule in which one or a few highly ordered heavy atoms are present. The native crystal and the derivative crystal should be isomorphous: the lattice parameters and macromolecular structure should be the same and the only difference should be the presence or absence of the heavy atoms. The structure factors of the derivative (Fph), the native (Fp) and of the heavy atom structure alone (Fh) are related by the relation $Fp = Fh + Fph$. The amplitudes $|Fph|$ and $|Fp|$ can be measured, and if Fh can be determined by direct methods, a vector diagram shows there are two possible solutions for the phase of Fp, of which only one is correct. This phase ambiguity can be resolved with a second derivative (or several more derivatives), hence the name MIR. The anomalous dispersion signal, see Section 1.3.5, can also solve the phase ambiguity; this technique is called single isomorphous replacement with anomalous signal or multiple isomorphous replacement using anomalous signal if more than one derivative is used.

In the absence of anomalous dispersion, structure factors follow Friedel’s law (Section 1.2.5). If the crystal contains atoms that resonate with the X-ray radiation, anomalous dispersion occurs, and Friedel’s law is no longer true. By comparing reflections that should be symmetrically related by Friedel’s law, an anomalous dispersion effect can be measured. The intensity signal due to the anomalous dispersion effect is small but can be optimized when the X-rays used are of the energy corresponding to the adsorption edge of the resonating atom. X-ray fluorescence emission scans indicate the magnitude of the effect and its dependence on wavelength.

The anomalous dispersion effect gives phase information and can be used in combination with isomorphous replacement as described above. In some cases, the SAD technique is sufficient to produce reliable phases. In other cases, MAD is necessary. This consists of measuring complete datasets from the same crystal at different wavelengths, usually three to five different wavelengths. For MAD or MAD to work, it is obviously necessary to have ordered heavy atoms (Se, Hg, Pt, Fe, Zn, Cu, etc.) in the crystal. A derivative that turned out not to be isomorphous and thus unsuitable for MIR may be used. It is also common to “label” the protein during expression with Se-Met. An advantage of the MAD and SAD techniques is that they do not have nonisomorphism problems, because the datasets are measured from the same crystal.

### 1.3.4.4 Density Modification

Once reasonable starting phases have been determined and the resulting maps show some interpretable features, density modification procedures can significantly improve them. These procedures use prior knowledge of the distribution of the electron density in the asymmetric unit. Solvent flattening uses the observation that in crystals of macromolecules a significant connected portion of the asymmetric unit is not occupied by the macromolecule but by solvent. We also usually know the size of our macromolecule, allowing us to make a reasonable guess for the solvent content. If the starting phases are good enough to estimate which parts of the asymmetric unit are occupied by solvent, the electron density in these regions can be set to a constant value, typically 0.33 e Å$^{-3}$ (protein electron density averages to 0.43 e Å$^{-3}$ but is not constant and shows strong local variation in the protein region). However, in the protein region, negative density should be absent, and this knowledge is also incorporated, by resetting to zero regions where the density is negative.

From previously solved protein structures, the expected density distribution in the protein region is also known. Using this density distribution histogram as a mold, small alterations are made to the protein density to make the experimental density distribution histogram match the expected one. If multiple copies of the macromolecules or complex are present in the asymmetric unit (i.e., noncrystallographic symmetry, NCS, is present) and envelopes can be identified for the NCS-related protomers, the density at NCS-related points within these monomers can also be averaged and imposed to be equivalent. The density modification process is cyclic: the modified map is back-transformed to give modified phases; these phases are recomputed with experimentally determined phases and a new map is calculated. This new map is then again modified. Programs for density modification include SOLOMON [58], DM [59], RESOLVE [60], and SHELXE [61].

### 1.3.4.5 Combined Methods

The methods mentioned above may be combined into procedures that are more powerful than any by themselves. Two examples of combining molecular replacement with direct methods are described here. The program ACORN locates small user-defined peptide fragments by molecular replacement and performs phase refinement by direct methods. It is useful for solving peptides and small proteins when high-resolution data are available (1.2 Å or better) [62]. The ARCIMBOLDO procedure also locates small model fragments, alpha-helices in this case. It uses PHASER for molecular replacement and performs
sophisticated density modification with SHELXE, allowing success for relatively large proteins and with data extending to 2 Å resolution or better [63].

Detailed discussions and explanations of macromolecular phasing methods are available in Taylor [64] and in several textbooks [65–69]. Programs and program pipelines commonly used for phasing include SHARP [70], SOLVE [60], SHELX [51], and CRANK [71].

1.3.5 Electron Density Map Interpretation: Model Construction

Once an interpretable electron density map has been obtained, a model for the protein will have to be built using molecular graphic programs or, if the map is of sufficient quality, in combination with automated building procedures (Fig. 1.4). An important quality measure is the resolution, which in general should be better than 2.3 Å in order to allow automated procedures to construct virtually complete models. However, the completeness of the dataset is also important, as is the necessity that all reflections, including low-resolution ones, are measured well. Maps calculated using data with missing or badly measured low-resolution reflections may suffer from reduced electron density connectivity and may be more difficult to use for constructing initial models. At intermediate resolutions (2.5–3 Å), automated structure building programs are unlikely to construct complete models but may still be useful for construction of parts of the model.

At lower resolutions, where a complete model needs to be constructed manually, the, for now, superior pattern recognition capability of the human brain plus additional knowledge about the macromolecule studied (expected fold, ligands, etc.) is employed. It is often useful to skeletonize the map.

FIGURE 1.4 Construction of protein models in electron density maps. Top left: electron density map obtained by experimental phasing using the program SOLVE. Top right: Map with skeleton calculated by COOT. Bottom left: Model obtained by autotracing with RESOLVE superimposed on the map. Bottom right: Refined protein model including some water atoms (yellow crosses). See insert for a color representation of the figure.
A skeleton is a collection of lines representing connected regions of electron density in the map. If the map is of sufficient quality, this skeleton should be similar to the protein chain trace and often will allow one to estimate the fold. The skeleton is edited to remove spurious connections and to introduce connections the procedure failed to identify. The program O [72] is useful for this and also contains other model construction facilities. Once the edited skeleton resembles the protein chain trace as much as possible, it is replaced bit by bit with amino acids, in the first instance by a polyalanine chain. Using the known protein sequence, regions of the map are then carefully inspected to see if short sequence of side chains can be recognized. Once the protein chain is reasonably complete and at least some of the side chains are identified, intermediate refinement runs (see Section 1.3.6) may be used to improve the density maps and calculate difference density maps although care has to be taken not to introduce “model bias” (model bias is caused by the calculated phases from the model biasing the resulting electron density to the model rather than to the measured diffraction intensities). The improved maps and difference density maps can then be used to identify nonmodeled density and wrongly modeled regions. A modern model construction program is COOT [73], which contains many tools for building, refinement, analysis and validation of protein and nucleic acid structures, ligands, and solvent molecules.

In high-quality electron density maps, automated model construction programs like ARP/wARP [74], RESOLVE [60], or BUCCANEER [75] can identify large fractions of the protein chain and solvent molecules. Some programs, like ARP/wARP, can also automatically build nucleic acid structures. The WARP/TRACE feature of ARP/wARP interprets electron density maps as free atom models, bonds atoms to each other if they are sufficiently close and resemble amino acids, joins the amino acids in a protein trace, and refines the resulting protein-solvent hybrid model. It used the hybrid model to calculate a new electron density map, which is then used to find new free atoms and remove atoms that no longer show electron density. It also docks the identified protein chain into the known sequence [76]. The RESOLVE automatic building procedure identifies helices and strands by matching templates to the electron density. Then, fragments of helices or strands from a library are matched to this density and extended in both directions using tripeptide fragment libraries. Subsequently, side chains are identified using libraries and the protein chains are assembled and are as compact unit as possible. BUCCANEER identifies likely C-alpha positions in the electron density map. The best-fitting ones are used as seed positions; the seeds are then grown using the other C-alpha positions in the map into extended chain fragments. The chain fragments are then docked into the known protein sequence.

When a complete protein chain has been built, difference maps can be used to identify ligands and solvent atoms. Ligands may be copurified from the expression system or added during purification and crystallization. Care should be taken not to overinterpret initial maps, because model bias may appear to confirm the presence of a ligand which may not really be there. Normally, many ordered water molecules will be observed on the surface of the crystallized macromolecule and can be validated both in terms of electron density and analysis of the hydrogen bonding with the macromolecule and other solvent atoms. Ordered glycerol from the cryoprotecting solution or precipitant molecules, for example, a partially ordered polyethylene glycol molecule, may also be observed. Chemical reactions may also occur in the crystallization drop, and their products may be observed, for example an oxidized dithiothreitol molecule. Some of the identified water molecules may in fact be ions, either metal ions or other ions such as ammonium, sulfate, carbonate, or phosphate. Careful inspection of the maps and analysis of the coordinating atoms will be necessary to identify these correctly. If the ligands or solvent molecules contain heavy atoms, they may show sufficient anomalous signal to help correct identification. The program COOT has incorporated functions to add many common ligand molecules for identifying and validating water atoms. Some automatic building programs, like ARP/wARP, also identify solvent atoms.

### 1.3.6 Model Refinement

Once a complete protein model, including ligands and ordered solvent molecules, has been built, the structure should be refined using appropriate geometric restraints and the best dataset available with respect to completeness and resolution. Refinement consists of making small changes to the positional parameters and temperature factors of all atoms simultaneously, using a certain target function. Traditionally, the target consisted of minimizing the $R$-factor (Equation 1.8), a factor that expresses residual disagreement between the observed structure factor amplitudes ($F_{obs}$) and the calculated ones ($F_{calc}$). The $R$-factor is still an important statistic quoted in articles reporting macromolecular structures. To avoid model bias, current practice is to remove a small fraction of reflections from the refinement target, which is then used to calculate the $R$-free value [77]. Upon refinement, the $R$-free should drop to a similar extent as the $R$-value, suggesting absence of model bias. As a general rule-of-thumb, an $R$-free value of less than 0.3 is considered reasonable, although this depends on the quality of the data and at higher resolution lower values are expected, while at resolutions worse than 3 Å, in some cases higher $R$-free values may be acceptable.

$$R = \frac{\sum |F_{obs}| - |F_{calc}|}{\sum |F_{obs}|}, \quad (1.8)$$

At the resolutions typical for data collected from macromolecular crystals, the total number of parameters to be refined is of the same order of magnitude as the structure factor amplitudes they are to be refined against. This low
The data-to-parameter ratio makes cross-validation necessary ($R$-free value, see previous paragraph) and imposes the use of restraints and constraints. Constraints effectively reduce the amount of parameters to be refined. At very low resolution, one may for instance refine groups of amino acids or whole protein domains as rigid bodies. At 1.5–3 Å resolution, refinement can only be performed if appropriate geometric restraints are included. These restraints effectively augment the number of data points and include the distances between bonded atoms, their bond angles, and certain torsion angles. The planarity of the atoms involved in peptide bonds, carboxyl and carboxamide groups, and aromatic groups is also restrained, as is the minimal distance between noninteracting atoms. Temperature factors may also be constrained to be the same between groups of atoms or restrained not to vary too much between neighboring atoms. NCS, if present, may also be used to constrain or restrain multiple copies of the same macromolecule to be the same or similar. In this case, one should keep in mind that legitimate differences may of course exist between certain portions, and these should be removed from the restraints. The most common programs used for refinement are REFMAC [78] and PHENIX [79]. REFMAC uses a maximum-likelihood target while PHENIX can also use a least-squares target.

Validation

Validation of the solved and refined macromolecular structure is a necessary quality-control step, as important errors in model building and refinement may have gone unnoticed. The validation process judges parameters used in refinement such as bond distances, bond angles, certain torsion angles, correctness of chiral centers, planarity of groups of atoms that show resonance (such as atoms involved in peptide bonds, carboxylate and carboxamide groups, and aromatic rings), van der Waals distances, hydrogen bonds, and coordination distances to metals. Usually, during refinement, care has already been taken to keep these parameters at sensible values, and they are therefore not truly independent parameters. The temperature factor distribution should also be sensible in that connected atoms should not have very different temperature factors and that high temperature factors are restricted to atoms that have room to move in the structure, that is, are on the surface of the protein. The validation programs may include these temperature factor restrictions.

Validation can also and should be used to verify independent parameters that were not used in refinement. A good example of parameters not usually refined are the phi and psi torsion angles, which are usually represented in a Ramachandran plot (Fig. 1.5) [80]. Certain combinations of phi and psi angles are much more common than others, while other combinations are highly unlikely or even physically impossible. The probability depends on the nature of the side chain of the amino acid. A glycine residue, which lacks a side chain, can adopt more different conformations than other amino acids, while proline, with its atypical side chain covalently bonded to both the alpha-carbon and the nitrogen atom, has a more restricted conformation space than other residues. A structure is also expected to contain all or nearly all of the peptides in trans-conformation, although in rare cases cis-peptides...
may occur, especially when the amino acid C-terminal to the peptide bond is a proline. Side chains also have energetically favored orientations (preferred rotamers), which can be expressed as combinations of their torsion angles (Chi1, Chi2, etc.). Certain amino acids have parameters that can be specifically checked. Prolines should show a distinct puckering, in which the gamma-carbon is rotated either above or below the approximate plane formed by the alpha-carbon, nitrogen, and other side-chain carbons. Asparagine, glutamine, and histidine residues have pseudosymmetric side chains; they should be positioned, and if necessary “flipped”, to optimize hydrogen bonding. The likely protonation state of the residues, especially histidine, should be taken account for this. Finally, an important validation parameter is whether all amino acids are in suitable environments in relation to their nature. Apolar and aromatic side chain should preferably be buried in the hydrophobic core of the protein. Polar groups should be in polar environments, either in contact with solvent or with other polar residues. Electrostatic charges should be neutralized by other charged atoms of the protein or solvent. Validation for nucleic acid structures is less developed and limited to checking of bond lengths and angles, sugar puckering, hydrogen bonding, and contact analyses.

Several computer programs are available for automated validation checks, either as stand-alone programs or as web servers. PROCHECK [81] checks basic validation parameters and outputs Ramachandran plots. WHATCHECK [82] performs more extensive checks. The web-based MOLPROBIT server [83] and the PHENIX [79] validation options are more modern implementations. POLYGON [84] compares model quality indicators to similar structures in the database.

1.4 STRUCTURAL ANALYSIS AND BIOLOGICAL IMPLICATIONS

Once the structure has been solved and preferably refined to completion, the structure will have to be analyzed. Firstly, to judge whether the structure is similar to other known structures or whether perhaps a new fold has been discovered. Further analysis concerns the biological interest of the structure, which in turn can provide hypotheses that can be tested by additional biochemical or structural analyses.

1.4.1 Structural Analysis

If the protein structure has been solved by molecular replacement, the final structure will have significant structural similarity to the input model and most likely will have the same fold. For *de novo* structure solutions, the program DALI can perform similarity searches against the protein structure database automatically [85]. The program outputs different similarity scores and a structural alignment, and also a superposition matrix. This matrix can be used to superimpose the structures and inspect them for structural similarity and differences using a structure visualization program. In the case of multidomain structures, analyses will have to be performed with all domains separately. In some cases, no structural homologs can be identified, and a truly new fold has been identified. However, in most cases one or more clear structural homologs can be identified. In any case, the topology of the new structure should be determined. The fold or topology is defined as the composition of secondary structure elements and their interconnection. If one or more structural homologs are identified, the topology of the new structure should be compared with the previously analyzed structures to see if they have identical topologies or whether there are some different interconnections between the secondary structure elements. The composition of secondary structure elements determines whether the protein falls in the family of alpha-helical structures, beta-structures, or mixed alpha/beta-folds. The SCOP (structural classification of proteins) [86] and CATH (class-architecture-topology-homologous superfamily) [87] databases both aim to further classify all the existing protein folds at different hierarchical levels. While some macromolecular structures exclusively exist as monomers, others form stable multimers, exist in different quaternary states depending on conditions, or are in dynamic equilibrium between different quaternary states. Due to the high concentrations used in crystalization, usually the highest possible multimeric state is observed in crystals. Furthermore, additional interaction interfaces are often observed in the crystal that turn out not to have biological relevance. The program PISA attempts to discriminate between genuine interaction interfaces and fortuitous crystal contacts by calculating interaction surfaces and complexation energies and entropies [88]. The quaternary structure in solution can also be investigated using analytical ultracentrifugation, dynamic light scattering techniques, or size exclusion chromatography.

1.4.2 Biological Implications

Although crystal structures are static, conclusions about protein movement can often be drawn. Local flexibility is often indicated by higher atomic displacement parameters of the corresponding atoms, although this is usually limited to loops on the surface of the protein. Comparison of different crystal forms of the same protein may locate hinges in the protein structure around which surrounding domains may move. Larger-scale movements can sometimes also be obvious. A clear example is F1-ATPase, where the presence of three alpha- and three beta-subunits alternating with each other in a ring around a seventh gamma subunit clearly indicated the possibility of relative rotation of the alpha-beta ring around gamma [89]. This rotation was later demonstrated, among other methods, by microscopy [90].
Inspection of the structure during modeling (Section 1.3.5), refinement (Section 1.3.6), and validation (Section 1.3.7) stages may have turned up ligand molecules associated with the macromolecule. These ligand molecules may have been copurified from the expression host, may have bound to the protein in the purification process, or may be components of the crystallization or cryoprotection mixture. Observed ligands may be substrates, products, cofactors, or inhibitors of the crystallized macromolecule in case it is an enzyme. In these cases, the observed structure and binding mode often allow reasonable proposals to be made about the reaction mechanism. Ligands may also mimic natural substrates, products, cofactors, or inhibitors (for instance, sulfate from the precipitant mimicking phosphate or multiple glycerol molecules mimicking a more complex carbohydrate).

Computational ligand docking may also be performed, either using the protein as a static entity or incorporating induced fit principles [91].

Analysis of a protein surface may indicate regions implicated in interaction with other biomolecules. Various structure-visualization programs may be used to predict surface potential, for example, GRASP [92], or either PYMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4. Schrodinger, LLC) or CHIMERA [93] in combination with APBS [94]. Proteins interacting with nucleic acids often have positively charged patches in the regions that interact with the RNA or DNA phosphate groups (for an example, see Reference 95). Shape complementarity may also be used to predict interaction partners; and computational protein-protein docking approaches may lead to useful binding mode hypotheses [96].

Experiments to verify hypotheses suggested by the crystal structure may be very diverse in nature. Binding assays may be performed to confirm interactions in solution and to measure binding affinity. Implication of specific amino acids or nucleotides in binding sites may be confirmed by site-directed mutagenesis and binding assays or other in vitro or in vivo experiments. New crystal structures may be determined with related ligands to test if their binding mode is similar to that observed in the original crystal. In general, the solution of the crystal structure of a new macromolecule opens up a multitude of new research directions, making the endeavor described in Section 1.3 very worthwhile.

1.5 FUTURE PROSPECTS

From its history of over half a century and the preceding paragraphs, it can be concluded that macromolecular crystallography is a mature technique, embedded in the mature science of structural biology. When well-diffracting crystals can be obtained, determination of small- and intermediate-size soluble protein structures is almost routine and seen as a technique among others to be used to understand a specific biological process. Indeed, in high-throughput structural proteomics projects, structures are determined in a matter of weeks from construction of the expression vector to a fully refined structure, although with a limited success rate (see for instance Reference 97). Automation of cloning, expression, purification, crystallization, X-ray data collection, and structure solution and refinement steps have significantly contributed to this. The development of better algorithms for phase determination has also been important.

Continuing improvements in X-ray beam intensity, combined with a reduction in beam size and larger and more sensitive detectors, also mean that ever larger macromolecular complexes can be studied (provided that they can be crystalized of course). A good example is the ribosome [98]. In spite of their importance for cellular processes and pharmacology, structures of integral membrane proteins are still under-represented in databases, but this is mainly due to the difficulty in producing them in large amounts and with the purity and homogeneity necessary for successful crystallization.

Protein structure prediction is another field in which significant improvements have been made and in several cases, predicted, albeit homologous, structures were good enough to serve as a molecular replacement model in de novo structure solution [99]. Nevertheless, in the near infinity of sequence and fold space there will always be “orphan” proteins with interesting functions and new folds which will need continued dedication of a specialized crystallography group in order to successfully determine their structures.

ACKNOWLEDGMENTS

We thank Carmela García-Doval and Bruno Dacunha-Marinho for careful reading of the manuscript. We acknowledge funding by the Spanish Ministry of Science and Innovation (grants BFU2008-01588 and BFU2011-24843) and the European Commission (BeNatural coordinated project, contract NMP4-CT-2006-033256).

REFERENCES

X-RAY CRYSTALLOGRAPHY OF BIOLOGICAL MACROMOLECULES: FUNDAMENTALS AND APPLICATIONS


15. Amann E, Ochs B, Abel KJ. Tightly regulated tac promoter vectors useful for the expression of unfused and fus...
REFERENCES

55. Ravelli RB, Leiros HK, Pan B, Caffrey M, McSweeney S. Specific radiation damage can be used to solve macromolecular crystal structures. Structure 2003;11:217–224.
22 X-RAY CRYSTALLOGRAPHY OF BIOLOGICAL MACROMOLECULES: FUNDAMENTALS AND APPLICATIONS


