Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Progress in Biophysics and Molecular Biology 89 (2005) 153–172
Progress in Biophysics & Molecular Biology
www.elsevier.com/locate/pbiomolbio

Review

Efficient use of synchrotron radiation for macromolecular diffraction data collection

Zbigniew Dauter

Brookhaven National Laboratory, Synchrotron Radiation Research Section, MCL, National Cancer Institute, Building 725A-X9, Upton NY 11973, USA

Abstract

In recent years, number of X-ray synchrotron beam lines dedicated to collecting diffraction data from macromolecular crystals has exceeded 50. Indeed, today most protein and nucleic acid crystal structures are solved and refined based on the synchrotron data. Collecting diffraction data on a synchrotron beam line involves many technical points, but it is not a mere technicality. Even though the available hardware and software have become more advanced and user-friendly, it is always beneficial if the experimenter is aware of the problems involved in the data collection process and can make informed decisions leading to the highest possible quality of the acquired diffraction data. Various factors, important for the success of data collection experiments and their relevance for different kinds of applications, are discussed.

Keywords: Synchrotron radiation; Diffraction data collection; Macromolecular crystals

Contents

E-mail address: dauter@anl.gov (Z. Dauter).

0079-6107/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.pbiomolbio.2004.09.004

1. Introduction

In the past few years macromolecular crystallography has undergone a period of very significant progress. Crystal structure elucidation has become more successful and less time-consuming to perform. A significant contribution to this trend has been the increased availability of synchrotron radiation sources. Since the beginning in the 1970s, synchrotron radiation has been highly valued, but the number of the available macromolecular synchrotron beam lines was limited. The increase in appreciation of the importance of X-ray structural investigations with the concomitant growth of the number of laboratories engaged in crystal structure investigations of macromolecules, and especially the advent of the high-throughput structural genomics initiatives, has persuaded various funding agencies to support the construction of several new synchrotron beam lines. At present, there are more than 50 active beam lines worldwide dedicated to macromolecular crystallography.

Synchrotron radiation has some unique properties, allowing its users to perform experiments not possible otherwise. The most important characteristics are the high intensity of the X-ray beam and the wavelength tunability. Outside of the mainstream applications are methods employing the white, non-monochromatized radiation and its pulse structure, which can be used in time-resolved structural investigations.

Synchrotrons are machines which accelerate charged particles to relativistic velocities and energies of the order of gigaelectronvolts. These particles, electrons or positrons, are then kept orbiting in high vacuum in the so-called ''storage rings''. The circular trajectory of particles in the ring is obtained with the use of bending magnets. In fact, storage ''rings'' consist of a number of bending magnets interspersed by straight sections. Since the orbiting particles passing through the bending magnet are exposed to a strong magnetic field and therefore to angular acceleration, they emit electromagnetic radiation in a very wide range, including hard X-rays. Moreover, the straight sections can be equipped with insertion devices, wigglers or undulators, which are multipole magnets, where particles undergo multiple kicks between alternating magnetic poles, producing an even more intense beam of r adiation than from the bending magnets. Fig. 1 illustrates the principles of these three types of radiation sources.

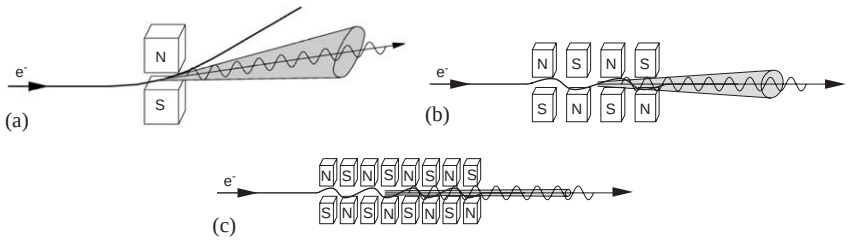


Fig. 1. Schematic representation of the three kinds of the synchrotron X-ray sources. (a) a bending magnet, with a single pair of magnetic poles, producing a fan of radiation relatively wide in the (horizontal) plane of the storage ring; (b) a wiggler, with several pairs of poles, giving much stronger flux of X-rays; (c) an undulator, where the positive interference of the generated X-rays enhance the intensity at a particular wavelength and its harmonic values, and produce a very narrow and highly parallel beam.

Collecting diffraction data from synchrotron facilities has become significantly easier in recent years. Several factors such as popularity of crystal cooling and transporting of frozen crystals, the advent of the automatic crystal-mounting robots, the routine use of the fast CCD detectors as well as the availability of fast and powerful computers, contributed to this progress. A very important role was played by the automation of the control of the beam line elements as well as of the process of data acquisition and reduction. Several semi-automatic systems, available at various synchrotron facilities, can be used to select optimal data collection parameters and to adjust the necessary beam line hardware components. Such systems are based on the theoretical principles of diffraction of X-rays as well as on the experience accumulated previously. However, since various macromolecular crystals differ enormously in their properties, it is rather difficult to develop an automatic system satisfying all possible scenarios. In less typical cases, the human brain remains indispensable for obtaining diffraction data of the best possible quality. Moreover, various requirements often contradict each other and a satisfactory compromise can only be achieved by a human operator.

Several texts discussing the ways of achieving good-quality diffraction data from macromolecular crystals are available in the literature (Dauter, 1997, 1999; Garman, 1999; Mitchell et al., 1999; Dauter and Wilson, 2001). In the following sections, the most important factors in the acquisition of diffraction data from macromolecular crystals at the contemporary synchrotron beam lines will be discussed. It is assumed that the screenless rotation method is employed (Arndt and Wonacott, 1977).

2. Requirements

The principal requirements for the success of a diffraction experiment are the accuracy of measured intensities and the completeness of the data set. Several factors have to be taken into account to achieve a complete and accurate data set, such as the characteristics of the crystal, the amount of time available for the experiment (usually severely limited), the practically achievable resolution of the data, the crystal resistance to radiation damage, and the purpose for which the data are to be applied.

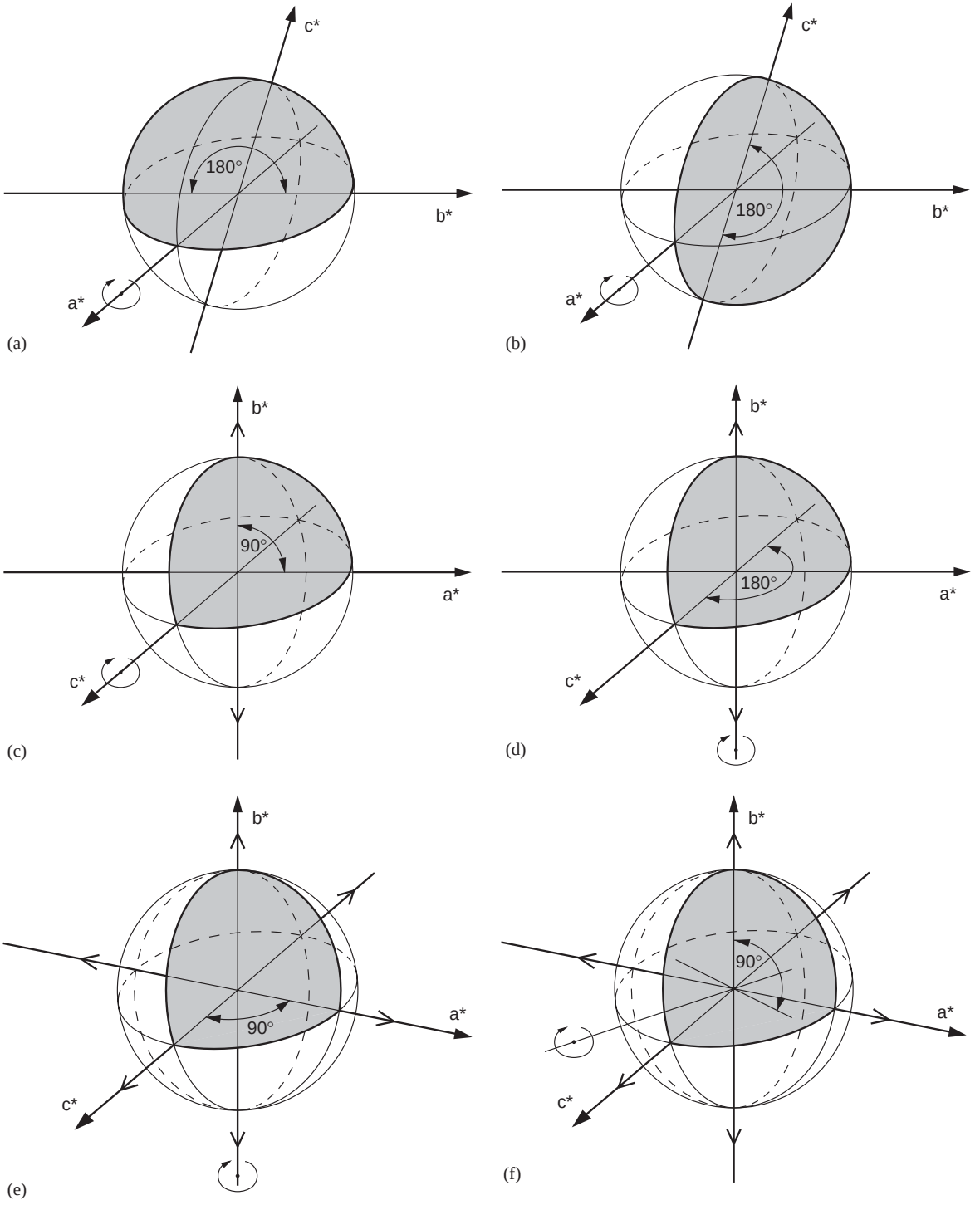
Different applications of the measured data may change the priorities among these various factors. If anomalous scattering data are collected, the highest priority should be the accuracy of measured intensities. If data are to be used for any method based on the Patterson function (e.g., for molecular replacement), they have to be complete at low-resolution with all strongest reflections recorded, since these methods rely on the squared amplitudes. For search of heavy atoms, data have to be complete, but do not need to extend to very high-resolution. For the refinement of the known model, the data should extend to the highest possible resolution. If the data are intended for the structure solution by direct methods at atomic resolution, the low-resolution reflections are important but it may also be advisable to liberally extend the high-resolution limit of diffraction and collect even very weak reflections. Obviously, in the ideal world it is good to fulfill all these requirements, but in practice it is virtually impossible and the experimenter must decide which combination of various parameters will lead to a satisfactory compromise.

The issue of data completeness has two aspects, quantitative, related to the reflection indices, and qualitative, related to the measured intensities. The first, quantitative context requires that all reflections within the complete asymmetric unit of the reciprocal lattice of the crystal are measured. This geometric completeness of the data set depends mainly on the cell dimensions of the crystal, its symmetry and on the orientation of the crystal with respect to the X-ray beam. The geometry and orientation of the detector are also important. The second, qualitative context relates to the collected intensities, which have to be accompanied by reliably estimated uncertainties of their measured values.

3. Completeness of the set of indices

Since X-rays are scattered from the atomic electrons, the final result of structural crystallography is the electron density map in the crystal unit cell. The atomic model of the investigated crystal structure represents only the chemical interpretation of the results of the diffraction experiment. Because of the Fourier relation between the direct (crystal) and reciprocal (diffraction) space, the electron density at each point in the crystal depends on contributions from all structure factors in the reciprocal space. Obviously, to obtain the accurate electron density map, all reflections have to participate in the Fourier synthesis; if some reflections are missing from the data set, the resulting map will be biased.

Fig. 2. The total amount of rotation necessary to collect the complete asymmetric unit of the native diffraction data for various crystal classes. For the anomalous data, it is necessary to cover two such neighboring regions. In the triclinic symmetry (a, b) it is always necessary to cover 180° of rotation, no matter where the starting point is. In the monoclinic symmetry, with its unique *b*-axis, if the spindle axis lies in the *a*, *c*-plane it is enough to cover only 90° of rotation between this plane and the unique axis (c); if a crystal is rotated around the *b*-axis, the full 180° of rotation is necessary (d), irrelevant of the starting point. In the orthorhombic symmetry, if the crystal is rotated around one of its twofold axes (e) or a vector in any of the principal planes (f), an appropriate 90° of rotation covers the complete asymmetric unit. A crystal in the class 4 rotated around its *c*-axis (g, h), requires 90° no matter where is then starting point, but if it is rotated around a vector in the *a*, *b*-plane (i) the specific 90° of rotation is necessary. A crystal in the class 422 needs 45° of rotation around its *c*-axis (j), but requires 90° if rotated around a vector lying in the *a*, *b*-plane (k).


c*
180°
b*
a*
(a)
c*
180°
b*
a*
(b)
b*
90°
a*
c*
(c)
b*
180°
a*
c*
(d)
b*
90°
a*
c*
(e)
b*
90°
a*
c*
(f)

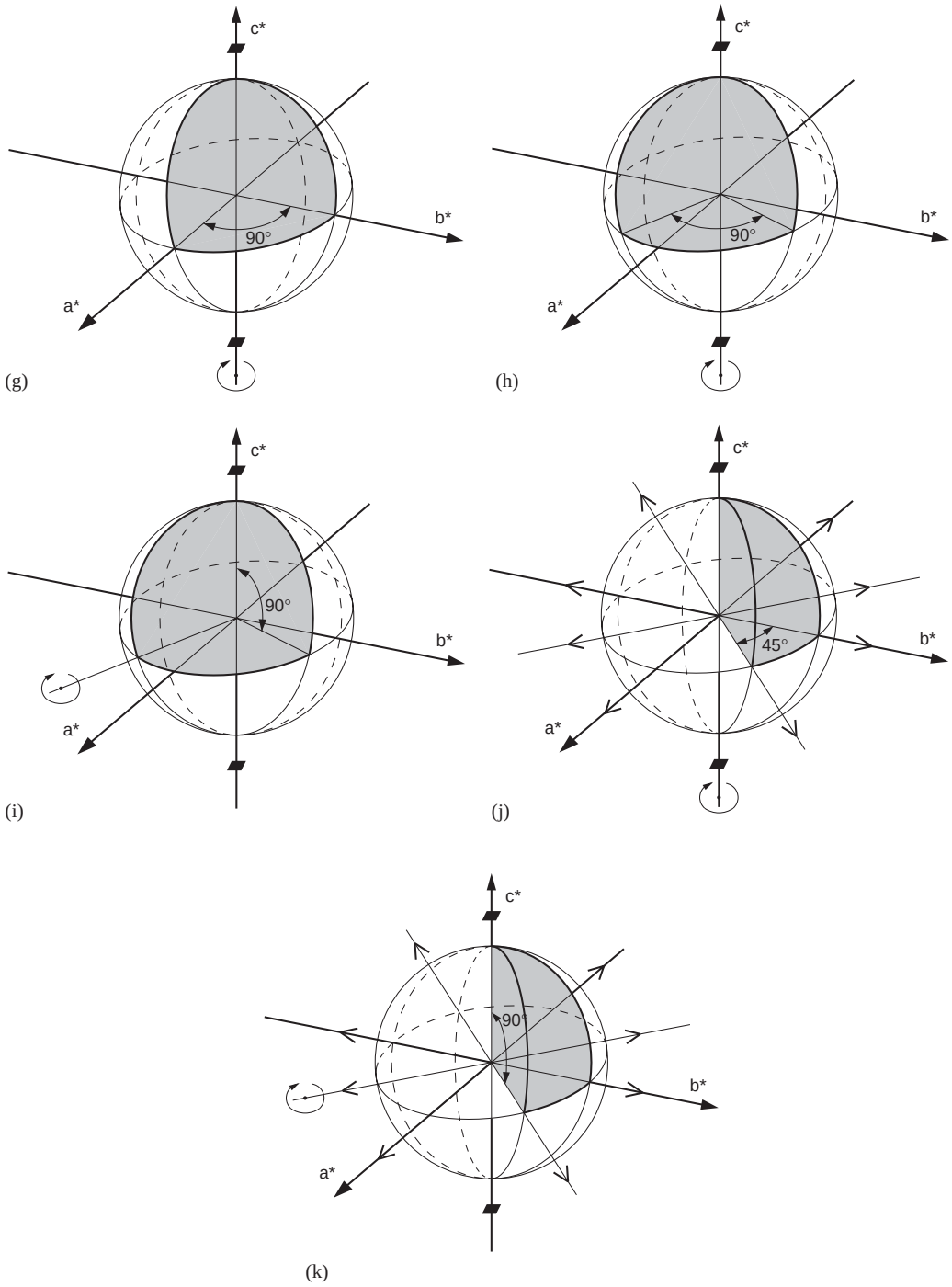


Fig. 2. (*Continued*)

If a small number of missing reflections were distributed randomly among all data, the effect on the electron density map would not be very detrimental. However, in practice, the missing reflections are always located systematically in the whole set. If less than the whole asymmetric unit has been collected, then reflections in a specific region of the reciprocal space are missing. If the dynamic range of the detector is limited, then a number of the strongest reflections will not be measured. In both cases, the electron density map will be strongly modulated.

The most important factor in achieving a full data set is the proper coverage of the complete asymmetric unit of the reciprocal lattice of the crystal. This depends on the crystal symmetry and its orientation in relation to the goniostat and the X-ray beam.

3.1. Coverage of the asymmetric unit

In contrast to the more familiar concept of the asymmetric unit in the crystal direct space, the asymmetric unit of the reciprocal space depends only on the point group (or more strictly, on the Laue symmetry group) of the crystal, and not on its space group. It always has the shape of a wedge with its apex at the origin and is bounded by the mirror planes of the corresponding Laue group. Away from the origin, it is limited by the maximum resolution sphere. Fig. 2 illustrates the possible definition of the asymmetric unit for different crystal symmetries. In the definition of the unique part of the reciprocal space, it is important to know whether the data are expected to contain the anomalous scattering signal, that is, if reflections related by the Friedel law should be treated as equivalent or not. In this context, it is advantageous to differentiate between the asymmetric unit, related to the native data, when any one of the Friedel-related reflections contributes to the overall completeness, and the anomalous unit, when both Friedel-related reflections need to be measured for a complete set. Obviously, the anomalous unit must contain two native asymmetric units related by the center of symmetry or by the mirror plane of the Laue group.

The minimum rotation range required to obtain a complete data set depends not only on the crystal symmetry, but also on its orientation with respect to the X-ray beam. Only for triclinic symmetry is the crystal orientation irrelevant, since in this case any hemisphere of the reciprocal space can be accepted as the asymmetric unit; and the whole sphere is required for the anomalous data. Fig. 2 illustrates that, for example, if the crystal of 422 class is rotated around its fourfold axis, the appropriate 45° rotation covers both asymmetric and anomalous units. However, if such a crystal is rotated around any of its twofold axes, the whole 90° is necessary. The monoclinic crystal requires 180° of rotation if mounted around its twofold axis, and 90° of rotation if mounted around any vector in the *a,c*-plane for the complete native data set, but for the anomalous data it always requires 180° of rotation, irrespective of how it is mounted.

These specifications relate to the "minimalist" approach; it is clear that collecting 360° of total rotation will always ensure a full completeness of data in any crystal symmetry (apart from the issue of blind region, discussed below). Collecting more than the minimal amount of data results in the higher multiplicity of measurements, but requires more time and involves the risk of the severe crystal radiation damage, inflicted by prolonged irradiation by X-rays. It is therefore advisable to start data collection at the proper crystal orientation, when its symmetry axes are either parallel or perpendicular to the X-ray beam. That ensures that the data completeness is reached as soon as possible, and extending the rotation range will increase the redundancy of

measurements. If radiation damage effects influence the later images, they can be rejected without the loss of completeness.

It is easy to visualize the required rotation ranges for various crystal classes (as in Fig. 2) if crystals are mounted around their axes of symmetry. It is virtually impossible to estimate the rotation range properly, if a crystal is oriented arbitrarily with respect to the goniostat spindle axis. Fortunately, the majority of data processing programs have "strategy" options, calculating the optimal rotation range and start point on the basis of one or a few initially interpreted diffraction images. It is highly advisable to make use of such options before starting a data-collection run at the beam line.

3.2. Blind region

In the rotation method, the crystal is rotated around a single spindle axis, usually perpendicular to the beam. Some reflections situated in the reciprocal space near this axis will never cross the surface of the Ewald sphere (Fig. 3a). Those reflections, located in the blind region, will never be collected, even after 360° rotation. Due to the curvature of the Ewald sphere, the width of the blind region (sometimes called "cusp") increases at high-resolution (Fig. 3b). It is narrower if the wavelength of radiation is short, and becomes wider at long wavelengths (Fig. 3c). The half-width of the blind region is equal to the diffraction angle at the highest resolution limit, θ_{max}.

If the crystal is triclinic, or if its unique axis (e.g., twofold in monoclinic symmetry or fourfold in tetragonal symmetry) is parallel to the spindle axis, all reflections in the blind region are lost and cannot be measured. The only remedy is then to reorient the crystal and collect more images covering the missing part of the reciprocal space. That can be done with the use of the kappa-goniostat.

However, if the unique axis of the crystal is oriented further away than θ_{max} from the spindle axis, all reflections within the blind region will have their symmetry equivalent mates located outside of the blind region and the overall completeness of the data will not be compromised (Fig. 3d).

The issue of the blind region is therefore relevant only if the crystal is triclinic or mounted along its unique axis, particularly when very high-resolution data are collected. In fact, it is impossible to collect atomic resolution data with long wavelength radiation; according to the Bragg law, $d = \lambda/(2 \sin \theta) > \lambda/2$, the maximum resolution cannot exceed half of the wavelength.

3.3. Beam divergence, crystal mosaicity and partial reflections

The finite sizes of the synchrotron X-ray source and of the beam collimators produce an X-ray beam that is not exactly parallel, possessing some angular spread. Moreover, if the beam is focused by the optical elements of the beam line, this adds to the overall divergence of the primary beam incident at the crystal sample. In this respect, it is again an example of another compromise, this time between the beam parallelness and its flux, since the more ideally the beam is collimated, the less flux ends up at the crystal. This situation is less acute at the newest third-generation synchrotrons, where the beam generated at the undulator sources is highly parallel and very intense, and often does not need to be strongly focused. At the bending magnet and wiggler sources, the fan of generated radiation is always wider in the horizontal plane than in the vertical

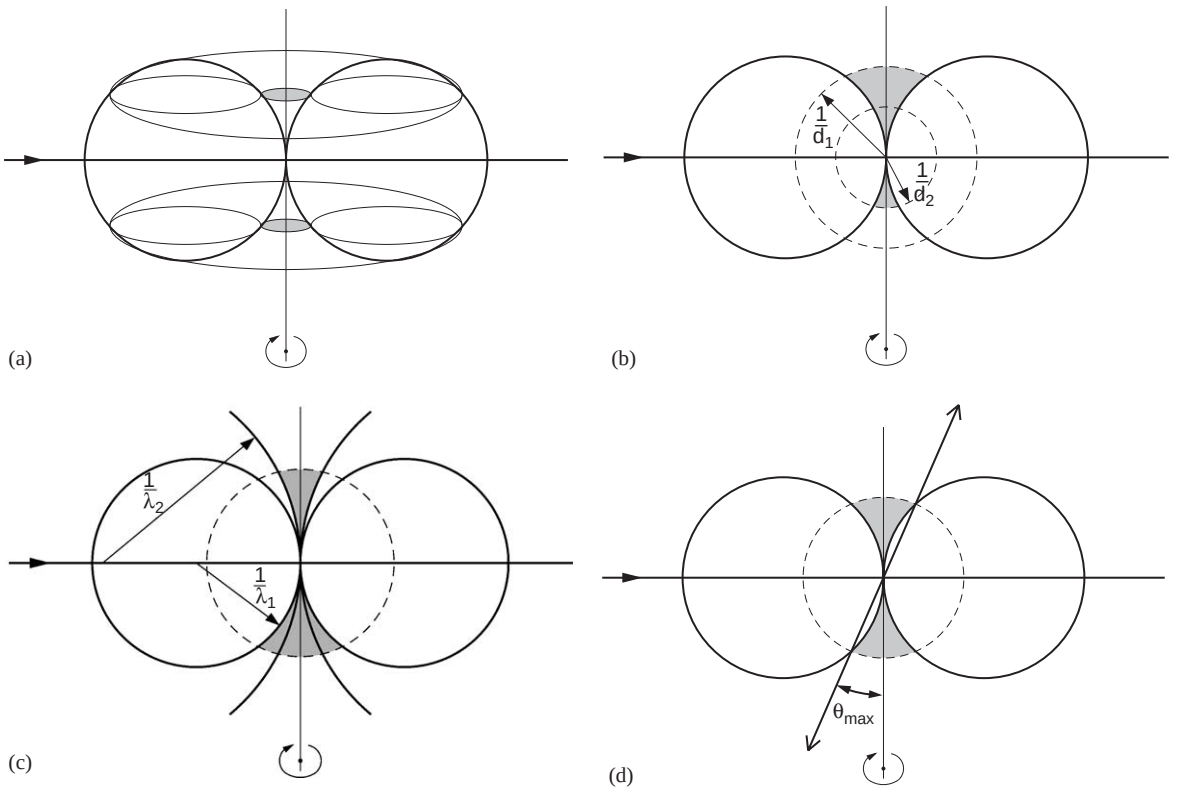


Fig. 3. Visualization of the blind region. The rotation of a crystal in the stationary beam is represented here as a rotating Ewald sphere in the stationary reciprocal space, which is geometrically equivalent. (a) Even after 360° of rotation in the blind region (marked gray) close to the spindle axis, will never cross the surface of the Ewald sphere. (b) Due to the curvature of the Ewald sphere, the width of the blind region is larger at high-resolution than at low-resolution. (c) The curvature of the Ewald sphere is larger for long wavelength, therefore the width of the blind region is smaller for short wavelength radiation. (d) If the unique symmetry axis is misoriented from the spindle by more than the θ angle at highest resolution, all reflections in the blind region have their symmetry equivalent mates outside of this region.

plane. That is the reason for the double-crystal monochromators being always set to rotate around the horizontal axis, which ensures narrower wavelength bandpass $\delta\lambda/\lambda$, important for the multiwavelength anomalous diffraction (MAD) experiments, exploring very fine features at the X-ray absorption edges of certain anomalously scattering elements.

According to the dynamic theory of diffraction, crystals are built from small mosaic blocks, which are slightly misoriented from each other. This affects the parallelness of the diffracted rays, and adds to the total width of the rocking curve of the crystal exposed to the X-ray beam (Fig. 4a). As a result, because of the beam divergence and crystal mosaicity, diffraction of each reflection is not instantaneous, but occurs over a finite time and a finite rotation range of the crystal during the data collection. In simplified form this can be represented in reciprocal space (Fig. 4b), by two limiting orientations of the Ewald sphere and by the finite size and a disk-like

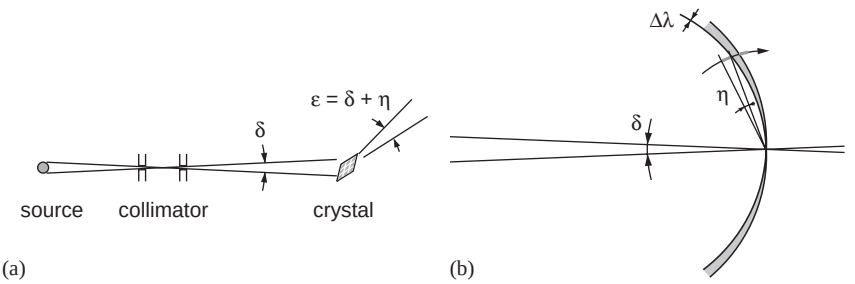


Fig. 4. Influence of the beam divergence δ caused by the finite size of the X-ray source and the collimator and of the crystal mosaicity η on the total rocking curve of the diffracted beams, represented schematically in the direct space (a) and in the reciprocal space (b).

shape of the reciprocal lattice point. The width of these points in radial direction is small, as required by the constant diffraction angle for a particular reflection.

Since diffraction data in rotation method are collected in the form of successive exposures corresponding to the successive small rotation ranges of the crystal, some reflections, which started diffraction during one exposure, will still diffract at the beginning of the next exposure. Consequently, various fractions of the total intensity of such ''partials'' are recorded on more than one image. In contrast, the fully recorded reflections are those for which all intensity is recorded at a single diffraction image.

Reflections in reciprocal lattices are grouped in families of parallel planes, and rays diffracted from reflections belonging to each plane form a cone, so that at the planar detector window such reflections are grouped in ellipses. Each family of parallel planes gives rise to a set of concentric ellipses (Fig. 5a). When the crystal rotates, reflections within each plane are grouped into lunes, limited by two extreme locations of each ellipse, corresponding to the beginning and the end of the rotation range within a single exposure (Fig. 5b). Partial reflections reside near the edges and fully recorded reflections near the center of each lune (Fig. 5c). If the rotation range per image is smaller than the crystal rocking curve, all reflections are partially recorded and there are no fully recorded ones.

Two approaches to the rotation method can be specified, depending on the relation between the rocking curve of the crystal and the rotation range. In the wide-slicing approach, the amount of rotation per exposure is comparable or larger than that of the rocking curve, whereas in the fine-slicing approach, the crystal is rotated by a small amount and the intensity of each reflection is spread over several images. In the fine-slicing mode, it is possible to build a three-dimensional profile for each reflection, in two dimensions of the planar detector window and in the direction of the spindle axis, whereas in the wide-slicing mode each profile is two-dimensional. The narrow rotation range results in a higher reflection-to-background ratio, whereas in the wide-slicing mode, background around each reflection profile accumulates even if the reflection does not diffract any more, resulting in the poorer signal-to-noise ratio. However, the fine-slicing mode is only beneficial if the detector used has a very short readout time; otherwise, the detector ''dead time'' makes the data collection process very inefficient. Various data reduction programs are more appropriate for either wide- or fine-slicing approaches.

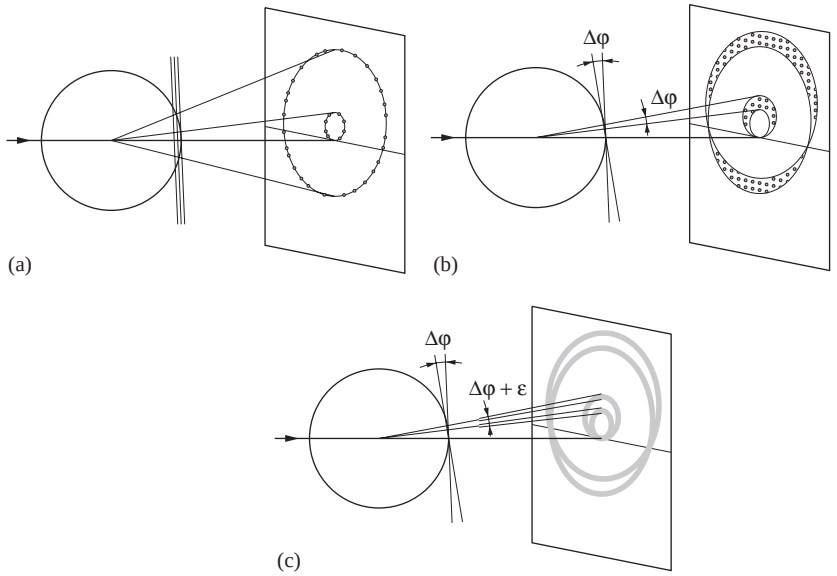


Fig. 5. (a) A family of reciprocal lattice planes gives rise to a set of cones of diffracted rays, and therefore to reflections forming the set of concentric ellipses at the detector. (b) If the crystal rotates, reflections at the detector are grouped in lunes, bounded by two positions of each ellipse, corresponding to the start and end of rotation. (c) Partial reflections, with intensity split between two successive images are grouped near the edges of each lune, and their number depends on the crystal mosaicity and beam divergence.

The gap between two consecutive lunes depends on the spacing within the family of the reciprocal lattice planes generating these lunes. The width of each lune in the direction perpendicular to the spindle axis is proportional to the amount of crystal rotation per image. If the rotation range increases, at a certain moment the two lunes at the highest resolution (at the edge of the detector window) will start overlapping (Fig. 6). This situation should be avoided, since then the individual reflection profiles from the two lunes may superimpose and it will not be possible to integrate properly their intensities. A simple formula can be used to estimate the maximum allowed amount of rotation per image:

$$\Delta\varphi_{\max} = (180d)/(\pi a) - \varepsilon,$$

where d is the limiting resolution, a the cell dimension along the primary beam direction (of the primitive unit cell) and ε the rocking curve. The most advisable approach for estimation of the rotation range in the wide-slicing mode is to first characterize the crystal and then to use the software strategy option, such as that provided by the program BEST (Popov and Bourenkov, 2003).

3.4. Non-equivalent indexing

In some crystal classes diffraction images can be indexed in more than one permissible, but not equivalent way. That may occur when the crystal point group symmetry is lower than the

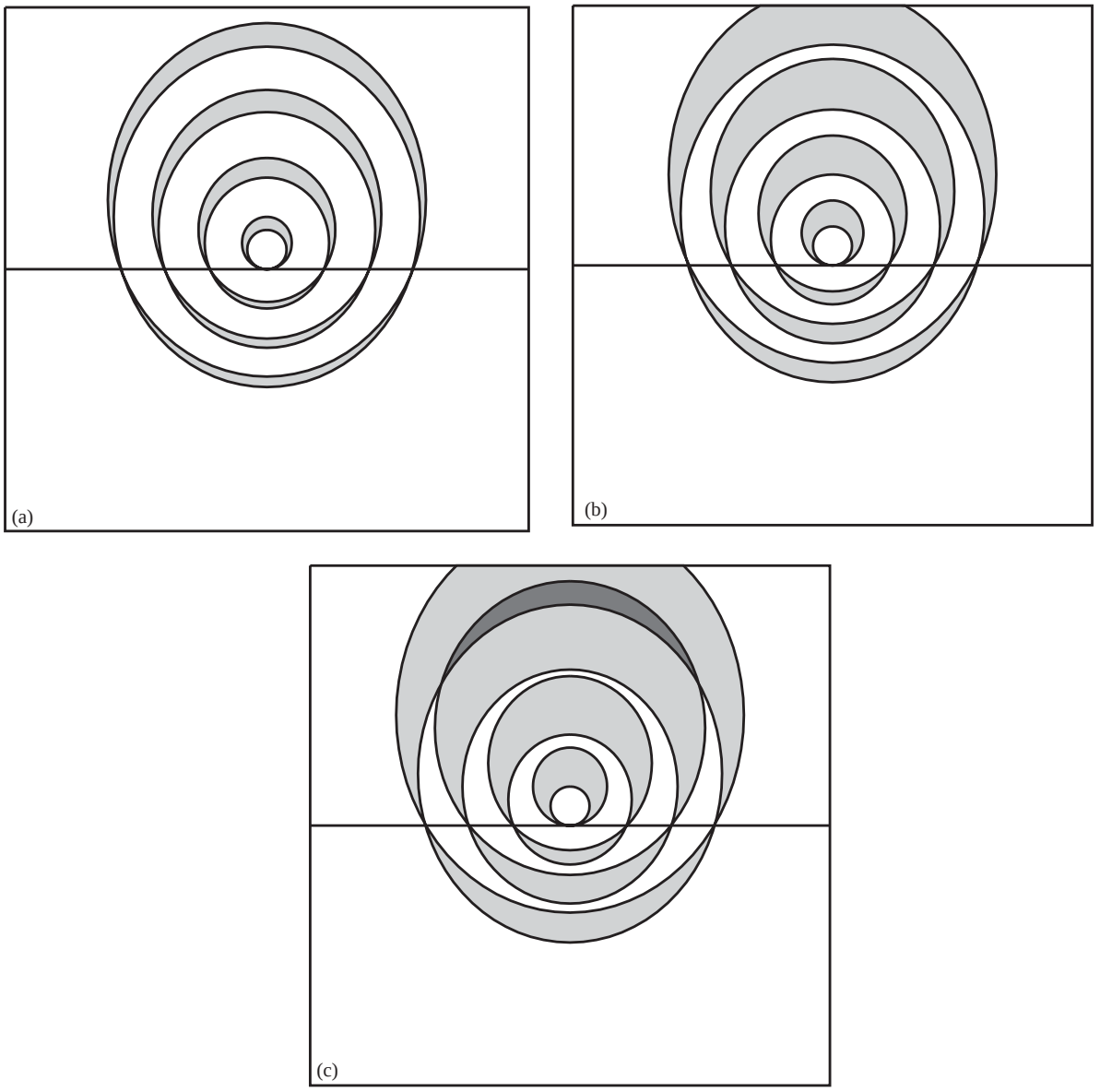


Fig. 6. The gaps between successive lunes are proportional to the spacing between corresponding reciprocal lattice planes and their width in the vertical direction (perpendicular to the spindle axis) depends on the amount of rotation per image (a, b). If the rotation range is large, two successive lunes start overlapping, and the individual reflection profiles may overlap as well and become not measurable (c).

symmetry of the crystal lattice. For example, in the tetragonal system, the lattice symmetry is at least 422 (in fact, it is 4/mmm, but since macromolecular crystals cannot have a center of symmetry, it is enough to consider a purely rotational group 422). In the crystal class 4, the

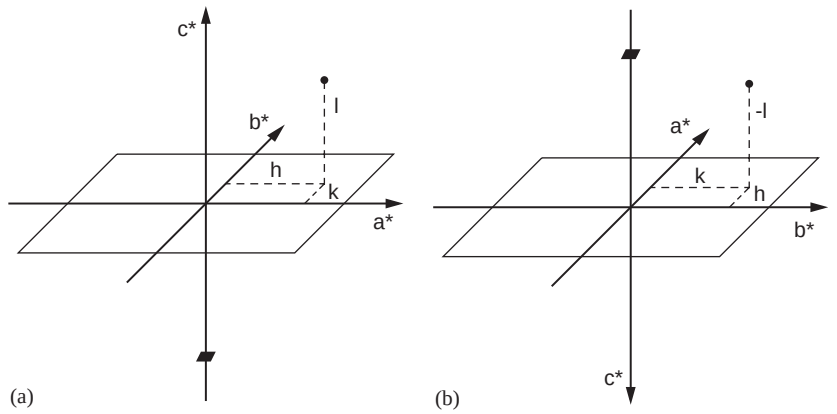


Fig. 7. In the crystal class 4 it is possible to index the diffraction pattern in two ways, represented by its fourfold axis pointing up (a) or down (b). Both ways are permitted but not equivalent, since a particular reflection (h, k, l) becomes (k, h, $-l$) which is not equivalent by the symmetry of this class.

fourfold axis is polar, and its two directions are not equivalent, whereas in the lattice symmetry 422 both directions of this axis are equivalent. As a consequence, a crystal in class 4 can be oriented in the tetragonal lattice in two ways, where the same reflection will have different indices, non-equivalent by the crystal symmetry (Fig. 7). Such a multiple choice of the reference frame can occur in the tetragonal class 4 (a subgroup of 422), hexagonal classes 3, 321, 312 and 6 (subgroups of 622) and in cubic class 23 (a subgroup of 432).

The same conditions lead to the potential occurrence of the merohedral twinning, when the crystal specimen consists of more than one domain, mutually oriented according to a symmetry operation existing in the lattice symmetry but absent from the true crystal point group symmetry. If the crystal cell parameters have special values, the metric of its lattice may correspond to a higher symmetry, e.g., for the monoclinic P cell with a unique β angle very close to 90°, the lattice is identical to that of the orthorhombic P cell. Such fortuitous agreement of the cell parameters may lead to the pseudo-merohedral twinning and a possibility of the alternative indexing of the diffraction pattern. The only characteristic that can be used to recognize the merohedral or pseudo-merohedral twinning is the statistics of the reflection intensities, and it is always advisable to check the cumulative intensity statistics of the measured data, particularly for crystals in the tetragonal, hexagonal and cubic systems.

The possibility of the alternative, non-equivalent indexing schemes in some space groups has to be taken into account when data are merged from more than one crystal or when derivative data are compared with a native set. In a MAD experiment, it is always good to use the same orientation matrix during integration of data from all wavelengths.

4. Completeness of intensities

Apart from the necessity to measure a complete set of reflections within the whole asymmetric (or anomalous) unit, it is important that all reflections have meaningful intensities accompanied

by their properly estimated uncertainties. In that sense, a data set with most of the highest resolution reflections lacking credibly measured intensities cannot be accepted as a complete data set to the nominally specified resolution limit. The judgment of data quality has to be based on the accuracy criteria of the measured intensities.

4.1. Quality criteria and limit of resolution

Several criteria are used in practice for the evaluation of the data quality. However, there is no single, global criterion, unanimously accepted in the community, since all of them have certain drawbacks.

The most widely used global quality indicator is the R_{merge}. It measures the spread of individual intensity measurements around the average value for the group of equivalent reflections. In its simplest and most popular version, it has the form

$$R_{\mathrm{merge}} = \left[\sum_{hkl}\sum_{i}|I_i(hkl) - \langle I(hkl)\rangle|\right] \Big/ \left[\sum_{hkl}\sum_{i} I_i(hkl)\right],$$

where $I_{\mathrm{i}}(hkl)$ are the individually estimated intensities of all reflections equivalent by the point group symmetry to $I(hkl)$, and $\langle I(hkl)\rangle$ is the average of all such intensities. It should be pointed out that R_{merge} in that form is not a statistically objective indicator, since it strongly depends on the data redundancy. Its value increases with increasing redundancy, when clearly the estimation of the average intensity becomes more accurate.

The more proper versions of an R factor have been proposed, such as R_{meas} (Diederichs and Karplus, 1997)

$$R_{\mathrm{meas}} = \left[\sum_{hkl} n\sum_{i}|I_i(hkl) - \langle I(hkl)\rangle|\right] \Big/ \left[\sum_{hkl}(n-1)\sum_{i} I_i(hkl)\right],$$

where n is a number of the symmetry equivalent contributors to the average, or a redundancy-independent version $R_{\mathrm{r.i.m.}}$ and a precision-indicating $R_{\mathrm{p.i.m.}}$ (Weiss and Hilgenfeld, 1997):

$$R_{\mathrm{r.i.m.}} = \left[\sum_{hkl}\sqrt{n}\sum_{i}|I_i(hkl) - \langle I(hkl)\rangle|\right] \Big/ \left[\sum_{hkl}\sqrt{(n-1)}\sum_{i} I_i(hkl)\right],$$

$$R_{\mathrm{p.i.m.}} = \left[\sum_{hkl}\sum_{i}|I_i(hkl) - \langle I(hkl)\rangle|\right] \Big/ \left[\sum_{hkl}\sqrt{(n-1)}\sum_{i} I_i(hkl)\right].$$

Unfortunately, these improved versions are still not widely used, probably because they have higher (albeit more informative) numerical values than the traditional R_{merge}.

The R_{merge} is vulnerable to various kinds of abuse. Obviously, reflections measured only once should not be included in the summation, since they would contribute only zeros to the numerator, artificially lowering the overall R_{merge} value. A small number of intensity measurements can be classified as outliers, which for some reason are influenced by significant and obvious errors. Such outliers may result from some detector pixels being inactive or wrongly calibrated. Zingers, characteristic for the CCD detectors equipped with the fiber glass taper systems, and resulting from the trace radioactivity of the taper glass, may influence some individual detector pixels. However, there should always be a reason why a particular intensity estimation is classified as an outlier. Obviously, outliers can be identified more reliably if the

redundancy of measurements is high. Too liberal rejection of outliers will artificially lower the value of R_{merge}. In extreme, R_{merge} can be made arbitrarily low, if enough measurements are rejected and the redundancy diminished towards unity.

Since R_{merge} depends on redundancy, it usually has a higher overall value for high-symmetry space groups than for triclinic or monoclinic crystals. It is usually lowest at low-resolution shells, where the measured reflections are strong, and increases towards the high-resolution limit of the data, where reflections are weak. As a rule of thumb, if R_{merge} at the highest resolution shell is lower than 25%, the data are most probably well measured and accurate; a value above 40–50% is normally not acceptable.

The ratio of the measured intensities to their estimated uncertainties, $I/\sigma(I)$, is from the statistical point of view a better quality indicator of the measured data, under the condition that the uncertainties are estimated confidently (see Section 4.2). Again, this indicator is more reliable if the data redundancy is high, since then both intensities and their standard deviations become more accurate. In principle, as long as the $I/\sigma(I)$ ratio is higher than unity, the data contain a certain amount of potentially useful signal. However, such data are usually too noisy for practical use. The commonly accepted resolution limit of the data usefulness is where $I/\sigma(I)$ drops below 2.0. To be able to judge the data strength, it is therefore necessary to know the overall value of $I/\sigma(I)$ as well as its value at the highest resolution range of the data set. The alternative criterion states that the limit of the useful data is at the resolution where 50% of reflections have intensities higher than $3\sigma(I)$. Since most of the contemporary phasing and refinement programs employ statistically valid weighting of reflection contributions, it is possible to be liberal with the definition of the data resolution limit. It is, however, important to properly estimate the uncertainties of measured intensities. The standard criteria of defining the data diffraction limit should be used in reporting the results in the form of publications and depositions.

All measured reflections should be included in the final data set, resulting from the data collection experiment. It is not a good practice to reject weak reflections or those with intensities estimated as negative. Due to statistical fluctuations, a small number of reflections are expected to have negative estimations. Their rejection inevitably introduces bias to all subsequent calculations. Their estimations can be modified on the basis of the Bayesian statistics, as implemented in the program *TRUNCATE* (French and Wilson, 1978) or they can be used directly in refinement against F^2, as is possible with *SHELXL* (Sheldrick and Schneider, 1997).

4.2. Uncertainties

As mentioned before, not only reflection intensities, but also their uncertainties should be estimated reliably in the data reduction procedure. According to the theory of counting statistics, the variance of the reflection intensity measured as the number of X-ray quanta is this number itself, or in other words, the standard deviation (accepted as an uncertainty) of the recorded intensity is the square root of the intensity. This is valid for detectors that measure individual X-ray quanta, such as scintillation counters used in four-circle diffractometers. However, the two-dimensional detectors based on CCD's and imaging plates do not measure the individual X-ray diffraction events, but reproduce numbers proportional to the diffracted X-rays, and the strict counting statistics are no longer valid. In practice, the estimated uncertainties have to be adjusted according to the statistically expected level. This can be done on the basis of the observed spread

of the intensities of symmetry equivalent reflections. Corrections are usually applied, in narrow resolution ranges, to make the uncertainties equal to the average difference of the individual intensity estimations from their mean value, $\langle I_i - \langle I \rangle \rangle$. High data redundancy is therefore beneficial for good estimation of uncertainties.

4.3. Amount of the anomalous signal

Several criteria can be used to estimate the amount of the anomalous signal in the measured data, such as

$$R_{\text{anom}} = \left[\sum_{hkl} \Delta I^{\pm}(hkl)\right] \Big/ \left[\sum_{hkl} I(hkl)\right]$$
$$= \left[\sum_{hkl} |I^{+}(hkl) - I^{-}(hkl)|\right] \Big/ \left[\sum_{hkl} I(hkl)\right]$$

$$\langle \Delta F^{\pm} \rangle / \langle F \rangle = \left[\sum_{hkl} |F^{+}(hkl) - F - (hkl)|\right] \Big/ \left[\sum_{hkl} F(hkl)\right]$$

or the anomalous signal to noise ratio, $\langle \Delta F^{\pm} \rangle / \langle \sigma(\Delta F^{\pm}) \rangle$.

The Bijvoet ratio of the anomalous signal to the total scattering signal, $\langle \Delta F^{\pm} \rangle / \langle F \rangle$, can be predicted if the number of anomalous scatterers N_{A} and their anomalous dispersion correction f''_{A} present among the total number of protein atoms N_{T} is known (Crick and Magdoff, 1956; Hendrickson and Ogata, 1997):

$$\langle \Delta F^{\pm} \rangle / \langle F \rangle = (2N_{\text{A}}/N_{\text{T}})^{1/2} (f''_{\text{A}}/6.7).$$

The observed values of $\langle \Delta F^{\pm} \rangle / \langle F \rangle$ are usually low at low-resolution and increase at high-resolution. In theory, f'' does not depend on the diffraction angle. However, the strong, low-resolution data are measured more accurately than the weaker high-resolution where the errors in the estimation of both numerator and denominator cause this ratio to increase towards unity. For well-measured anomalous diffraction data, the $\langle \Delta F^{\pm} \rangle / \langle F \rangle$ ratio tends to the calculated value at low-resolution (Dauter et al., 2002).

Sheldrick proposed a reliable and practical way of judging the quality of anomalous data in the form of a correlation between the signed anomalous differences in two data sets collected from the same crystal (Schneider and Sheldrick, 2002). These two sets can be measured at different wavelengths, as in MAD experiment, or merged as two partial sets from a more redundant single set. A meaningful anomalous signal extends to the resolution where the correlation coefficient is still higher than 30%.

4.4. Overloads

All two-dimensional detectors have a practical limit of the maximum intensity that can be reliably measured and read-out. In the CCD and imaging plate detectors, this limit results from the number of bits supported by the electronic read-out system of the detector. In most detectors, the maximum number stored in each pixel is 2^{16}–1, so that any intensity higher than 65 535 cannot be stored and such a pixel becomes electronically "overloaded". If a profile of a reflection contains overloaded pixels, its intensity cannot be measured properly, and such a reflection is absent from

the data set. The overloaded reflections have "top hat" profiles. If only a few pixels in the reflection profile are overloaded, it is possible to extrapolate their intensities by overlapping the standard reflection profile shape on the pixels in the shoulders of the complete profile and recover the total reflection intensity. Such a procedure may be advisable for certain applications (e.g., molecular replacement), but it does not provide a proper estimation of the highly overloaded reflections.

The issue of overloads is another example of a systematic, non-random lack of data completeness, since the overloads are the most intense reflections, present in the low-resolution region of the data. These reflections very strongly modulate the Fourier maps, and lack of them would inevitably degrade the electron density maps, causing the appearance of some spurious features and making their interpretation more difficult. These reflections play a particularly important role in all methods based on the Patterson function, where reflection amplitudes are squared. A small percentage of overloads may cause a total failure of a structure solution by molecular replacement methods. Similarly, the presence of all strongest, low-resolution reflections is very important for success of the direct methods solution of the heavy-atom substructures or atomic resolution protein structures.

It is therefore very important that the strongest reflections are measured properly in the synchrotron data collection experiment. If collection of the weak high-resolution reflections requires long exposures, it is necessary to measure the strong reflections in a separate data collection pass. This "low-resolution" pass should cover only that resolution range where the overloads appear in the long exposure images and therefore the rotation range per image and the crystal-to-detector distance can be increased. Obviously, the exposure time (or the X-ray beam attenuation) should be adjusted to minimize the number of overloads. For successful scaling of both high- and low-resolution sets, the effective exposure should not differ more than tenfold; otherwise the third, intermediate data collection pass may be necessary. In practice, it is advisable to collect the low-resolution pass first, followed by the high-resolution pass; otherwise the most important, strongest data are collected from a crystal already exposed to significant radiation damage. The short-exposure, low-resolution data collection pass does not cause such pronounced damage and should be performed first.

In the context of overloads, it would be advisable to include in the data quality reports not only the overall and highest resolution completeness of the data, but also completeness at the lowest resolution range.

4.5. Radiation damage

Radiation damage haunted the protein crystallography from its earliest days, when data were collected from crystals sealed in glass or quartz capillaries and kept at ambient temperatures. In such conditions it was rarely possible to obtain a complete data set from one specimen before radiation damage significantly degraded the crystalline order and impaired the crystal diffraction properties. Even after introduction of the crystal cryo-cooling as a routine practice, the deleterious effects of radiation damage could be very pronounced, especially at the strong synchrotron beam lines. Most crystals can withstand no more than 5–10 min of total exposure to a full beam from the undulator sources at the third-generation synchrotrons. Because exposure times at such stations are in the order of seconds, it is possible to acquire complete data rapidly, but the

influence of radiation damage has to be taken into account in the planning and execution of the experiment.

Irradiation of protein crystals causes damage through primary events, such as the breakage of interatomic bonds as a direct result of the absorption of a highly energetic X-ray quanta, and through secondary events, resulting from the propagation of radicals created by the absorption events. The secondary damage can be diminished by crystal cooling, but the primary damage does not depend on the sample temperature. In general, as a result of irradiation, at first the ordered atoms become disordered but still contribute to the crystal diffraction; later parts of the sample become amorphous and the sample loses its diffraction properties (Blake and Phillips, 1962; Hendrickson, 1976). The global effect is therefore a loss of diffraction by the crystalline sample. However, the first consequences of the radiation damage can be observed as the specific structural changes, such as the breakage of disulfide bridges (Weik et al., 2000), removal of bromine atoms from the Br-substituted bases in DNA (Ravelli et al., 2003) and many lesser changes.

Radiation damage and the induced structural changes influence the reflection intensities. It has been proposed that this effect can be neutralized by the interpolation of the intensities to zero-damage state (Diederichs et al., 2003), if the data redundancy is high enough. The effects of radiation damage can even be used to solve the crystal structures by the radiation-damage induced phasing (RIP) approach, in a fashion analogous to the SIR method, if multiple data sets with a variable degree of the structural damage are available (Ravelli et al., 2003; Banumathi et al., 2004).

Radiation damage can seriously impair the MAD phasing procedures, where multiple data sets have to be collected from the same crystal. It can cause a severe non-isomorphism between data sets collected first and last in the MAD series. In such cases, it may be often more beneficial to resort to the single-wavelength anomalous diffraction (SAD) phasing (Rice et al., 2000).

Experimenters performing synchrotron data collection on macromolecular crystals should be always aware of the radiation damage since, if not taken into account, it can considerably degrade the data quality and impair the process of phasing and model refinement.

5. Purpose of data collection

Various data quality factors may have different importances and priorities, depending on the purpose for which the data are to be applied.

Native data for the ultimate structure refinement should extend to as high a resolution as possible. For the purpose of the model refinement, the data do not need to be ultimately complete in the highest resolution ranges, but the low-resolution data have to be complete; otherwise, the fine features in the electron density maps would be biased and not reliably interpretable.

The data intended for structure solution by the MAD, SAD or MIR techniques do not need to extend to the maximum diffraction potential of the crystal. The highest priority should be directed towards the data accuracy and completeness of the strongest, low-resolution reflections. It is often more beneficial to solve the novel structure first, using complete and high-quality data at modest resolution and without inflicting too much radiation damage, and subsequently refine the so obtained model against the separate, high-resolution data set, than to attempt to collect the ultimate data for both structure solution and ultimate refinement.

If the structure is to be solved by molecular replacement, the data have to be complete at low-resolution, since only relatively low-resolution reflections are used for that purpose. This method relies on the Patterson function calculated from squared amplitudes; therefore all of the strongest reflections should be present in the data. If some overloaded reflections cannot be avoided, it may be advisable to include their approximately estimated values even at the cost of degraded accuracy.

If the atomic (beyond 1.2 A resolution) data are collected for structure solution by direct methods, the quality criteria can be relaxed to some extent. Even if at the highest resolution range the R_{merge} and $I/\sigma(I)$ exceed the normally accepted limits, it may contain a number of reflections with significant intensities that will facilitate the direct methods phasing process. This advice is applicable only to the ab initio phasing attempts, not to the solution of the heavy or anomalous atom substructure.

It is impossible to formulate the general synchrotron data collection protocol valid for all possible scenarios. Various applications and widely different crystal properties necessitate that the data collection protocols be appropriately adapted for each case, and usually it is necessary to find an optimal compromise between various, sometime contradictory requirements.

References

Arndt, U.W., Wonacott, A.J., 1977. The Rotation Method in Crystallography. North Holland, Amsterdam.

Banumathi, S., Zwart, P.H., Ramagopal, U.A., Dauter, M., Dauter, Z., 2004. Structural effects of radiation damage and its potential for phasing. Acta Crystallogr. D 60, 1085–1093.

Blake, C.C.F., Phillips, D.C., 1962. Biological effects of ionizing radiation at the molecular level. In: Proceedings of symposium held by the International Atomic Energy Agency, Brno, Czechoslovakia, 2–6 July 1962. IAEA, Vienna. pp. 183–191.

Crick, F.H.C., Magdoff, B.S., 1956. The theory of the method of isomorphous replacement for protein crystals. Acta Crystallogr 9, 901–908.

Dauter, Z., 1997. Data collection strategy. Method Enzymol. 276, 326–344.

Dauter, Z., 1999. Data-collection strategies. Acta Crystallogr. D 55, 1703–1717.

Dauter, Z., Wilson, K.S., 2001. Principles of monochromatic data collection. International Tables of Crystallography. vol. F. Kluwer Academic Publishers, Dordrecht, pp. 177–195.

Dauter, Z., Dauter, M., Dodson, E., 2002. Jolly SAD. Acta Crystallogr. D 58, 494–506.

Diederichs, K., Karplus, P.A., 1997. Improved *R*-factor for diffraction data analysis in macromolecular crystallography. Nat. Struct. Biol. 4, 269–275.

Diederichs, K., McSweeney, S., Ravelli, R.G.B., 2003. Zero-dose extrapolation as part of macromolecular synchrotron data reduction. Acta Crystallogr. D 59, 903–909.

French, G.S., Wilson, K.S., 1978. On the treatment of negative intensity observations. Acta Crystallogr. A 34, 517–525.

Garman, E., 1999. Cool data: quantity and quality. Acta Crystallogr. D 55, 1641–1653.

Hendrickson, W.A., 1976. Radiation damage in protein crystallography. J. Mol. Biol. 106, 889–893.

Hendrickson, W.A., Ogata, C.M., 1997. Phase determination from multiwavelength anomalous diffraction measurements. Method Enzymol. 276, 494–523.

Mitchell, E., Kuhn, P., Garman, E., 1999. Demystifying the synchrotron trip: a first time user guide. Structure 7, R111–R121.

Popov, A.N., Bourenkov, G.P., 2003. Choice of data-collection parameters based on statistic modelling. Acta Crystallogr. D 59, 1145–1153.

Ravelli, R.G.B., Schrøder-Leiros, H.K., Pan, B., Caffrey, M., McSweeney, S., 2003. Specific radiation damage can be used to solve macromolecular crystal structures. Structure 11, 217–224.

Rice, L.M., Earnest, T.N., Bunger, A.T., 2000. Single-wavelength anomalous phasing revisited. Acta Crystallogr. D 56, 1413–1420.

Schneider, T.R., Sheldrick, G.M., 2002. Substructure solution with SHELXD. Acta Crystallogr. D 58, 1772–1779.

Sheldrick, G.M., Schneider, T.R., 1997. SHELXL: high-resolution refinement. Method Enzymol. 277, 319–343.

Weik, M., Ravelli, R.B.G., Kryger, G., McSweeney, S., Raves, M.L., Harel, M., Gros, P., Silman, I., Kroon, J., Sussman, J.L., 2000. Specific chemical and structural damage to proteins produced by synchrotron radiation. Proc. Natl. Acad. Sci. USA 97, 623–628.

Weiss, M.S., Hilgenfeld, R., 1997. On the use of merging R factor as a quality indicator for X-ray data. J. Appl. Crystallogr. 30, 203–205.