Chapter 17

Data Collection for Crystallographic Structure Determination

Kanagalaghatta Rajashankar and Zbigniew Dauter

Abstract

Diffraction data measurement is the final experimental step of crystal structure analysis; all subsequent stages are computational. Good-quality data, optimized for a particular application, make the structure solution and refinement easier and enhance the accuracy of the final models. This chapter describes the principles of the rotation method of data collection and discusses various scenarios that are useful for different types of applications, such as anomalous phasing, molecular replacement, ligand identification, etc. Some typical problems encountered in practice are also discussed.

Key words Diffraction data collection, Diffraction data quality, Rotation method, Strategy

1 Introduction

Any X-ray structure determination project involves several steps, including selecting a target, cloning the gene, expressing the gene to obtain a sufficient amount of the protein, crystallizing the protein, collecting the diffraction data, and determining and refining the structure. Clearly, diffraction data collection is the last truly experimental step of the X-ray structure solution process. All subsequent stages are computational and can easily be repeated with different programs, algorithms, and parameters. Good-quality data make all of the computations easier and the resulting structural model more accurate. It is therefore important to carefully fine-tune all data collection parameters in order to obtain a diffraction data set best suited to the particular application.

Three important (and somewhat mutually contradicting) characteristics of an ideal data set are its completeness, resolution, and redundancy (in that order). First, the diffraction data set should be complete, i.e., it should contain all possible unique reflections, and all recorded intensities should be measured accurately and be accompanied by reliably estimated uncertainties. Second, the data should extend to the highest possible resolution. However, aiming at

Wayne F. Anderson (ed.), *Structural Genomics and Drug Discovery: Methods and Protocols*, Methods in Molecular Biology, vol. 1140, DOI 10.1007/978-1-4939-0354-2_17, © Springer Science+Business Media New York 2014

a very high resolution requires a higher X-ray dose/longer exposure, which may result in radiation damage and, in turn, incomplete data. Third, the data should be redundant, to provide good counting statistics. However, aiming at very high redundancy may result in lower-resolution data. Hence, one has to perform a “balancing act” to obtain the best possible data for the goals of the project and the potential of the crystal in hand.

In practice, the measured data are never ideal, but are usually influenced to some extent by various errors and discrepancies. Crystallographic experiments with different goals require different characteristics of the measured diffraction data [1, 2]. For example, phasing of a novel protein structure using an anomalous diffraction approach requires data that are different from those required for phasing by molecular replacement or for the ultimate, high-resolution refinement of the atomic model. Hence, one has to fine-tune the data collection parameters to suit the goal of the experiment. In the following discussion, data quality requirements for different types of experiments will be described.

1.1 Single- or Multi-Wavelength Anomalous Diffraction (SAD/ MAD)

The methods based on anomalous diffraction require the utmost accuracy of the measured intensities because they aim to utilize the inherently small anomalous signal for phasing. The data resolution does not need to extend to the full diffraction capability of the crystal. In fact, it may be more advantageous if one is not excessively ambitious, and limits the effective exposure, by utilizing less than the full diffraction potential of the sample. Aiming at the highest-resolution data results in excessive radiation damage, which can significantly degrade the accuracy of the data. However, the data set should be complete at low resolution, with all strongest, low-resolution reflections measured accurately.

1.2 Single or Multiple Isomorphous Replacement (SIR/ MIR)

The data quality requirements of these methods are similar to those of SAD/MAD, although perhaps not so stringent. These requirements are more relaxed because the isomorphous signal is usually stronger than the anomalous signal. The effects of radiation damage should be avoided, and the exposure times should not be selected overzealously. Of course, since, in this case, more than one set of data is measured from several crystals, the possibility of non-isomorphism between different crystals has to be taken into account.

1.3 Molecular Replacement (MR)

The data meant for molecular replacement do not need to extend to high angles, since, in this type of calculation, only relatively low-resolution data are used. Again, all strong, low-resolution reflections should be measured completely, since they play an especially important role in this approach, which is based on the Patterson function. Omission of these reflections is equivalent to setting their intensities to zero, which would severely bias the calculated Patterson synthesis.

1.4 Direct Methods

If one is collecting data to determine the structure using direct methods, one has to aim for atomic resolution, i.e., 1.2 Å or higher. To attain such data resolution, the crystal must be subjected to a high X-ray dose. In such situations, the strong, low-resolution reflections reach the saturation level of the detector. To overcome this problem, the data should be collected in multiple passes: first, a low-resolution pass with a lower X-ray dose to accurately measure the low-resolution data, followed by a high-resolution pass with a higher X-ray dose to obtain reasonable intensity counts for the weak, highest-resolution reflections. Data completeness at low-resolution is also very important.

1.5 Refinement

The diffraction data intended for the ultimate model refinement should extend to as high a resolution as the crystal is reasonably able to provide. A certain amount of radiation damage is then unavoidable, but it should not be excessive, so that the crystal should not "die" before the completion of the data set. A small amount of missing data is acceptable, but it should be remembered that missing reflections (especially the strongest) always deteriorate the appearance of all Fourier maps, biasing the interpretation of fine structural features. For very well-diffracting crystals, it may be advisable to use multiple passes for data collection.

1.6 Ligand Finding

The highest priority in experiments intended for the initial search for potential ligands is rapid turnover. Because ligand identification is usually based on difference Fourier maps, the total data completeness and resolution are not so crucial. For the proper structural analysis, data from such initially identified complexes may be more accurately and comprehensively measured later.

1.7 The Reality

In practice, however, diffraction experiments are often performed only once, and the same data are used for structure solution using the SAD, MAD, or MR methods and for the final model refinement. This is especially relevant in structural genomics applications, in which some compromises are unavoidable. Seeking too ambitious a resolution limit may result in a data set that is only partially complete and significantly deteriorated due to radiation damage. At the same time, too much attention to data accuracy will not deliver the highest possible resolution or the required data completeness. Selection of the appropriate protocols and fulfilling most of the requirements, without excessively degrading any single one of them, requires careful adjustment of various experimental parameters. It is not always easy, even for experienced experimenters, to optimally select these parameters.

Since the human eye is a very good detector of patterns, it is good practice to visually inspect one or a couple of the initial, test diffraction images and also to index them and check the predicted diffraction patterns. It is very easy to see whether the diffraction image shows a single lattice or comes from multiple crystals, whether the reflection profiles are acceptable and excessively smeared or overlapping, or whether there are too many overloaded pixels. In addition, it is important to check that the beamstop is positioned correctly and does not allow the direct beam to go through or create too much scatter around the detector center or, conversely, cast a shadow that is unnecessarily large, obstructing some of the low-resolution reflections.

Before proceeding with collecting the data, it is beneficial to take the time to run one of the available data collection strategy programs [3–6]. Using one or two initial test diffraction images, these programs are able to provide a set of optimized parameters based on the realistic estimation of the crystal, beam, and detector characteristics.

Another approach employs advanced robotization of the data collection process, which involves recording test exposures from a number of automatically positioned and centered crystals of equivalent specimens, indexing and integrating the initial diffraction images, and subsequently prioritizing all crystals on the basis of the obtained statistics. The best-diffracting crystal is then used for collecting the full data set. Such an approach is becoming increasingly popular in high-throughput structural biology, as it can save human effort and increase the experimental throughput.

2 Geometrical Principles of the Rotation Method

In the rotation method [7], diffraction data are recorded on a (usually) flat detector positioned in front of the crystal, whereas the X-ray beam is delivered perpendicular to the detector. If the X-rays are monochromatic, with the wavelength λ, only reflections fulfilling Bragg's law, $\lambda = 2d\,\sin\theta$, give rise to diffraction, which can be illustrated by the Ewald construction (Fig. 1). To bring more reflections to diffraction, the crystal is rotated by small angular amounts during exposure (Fig. 2). If the goniostat is equipped with additional axes, these are used only to set the crystal in the appropriate orientation; however, during data collection, the crystal is always rotated around a single axis (usually referred to as the omega axis), perpendicular to the incoming X-ray beam. Proper selection of data collection parameters will result in optimal data. The good news is that only a few parameters need to be optimized, namely, crystal-to-detector distance (affects the maximum data resolution); total and per-image rotation range (affects data completeness and the spot separation in a diffraction image); exposure time and X-ray flux (these two affect the diffraction strength); and choice of wavelength (applicable mainly for SAD/MAD experiments). The following sections describe these parameters and explain how they affect the data collection procedure.

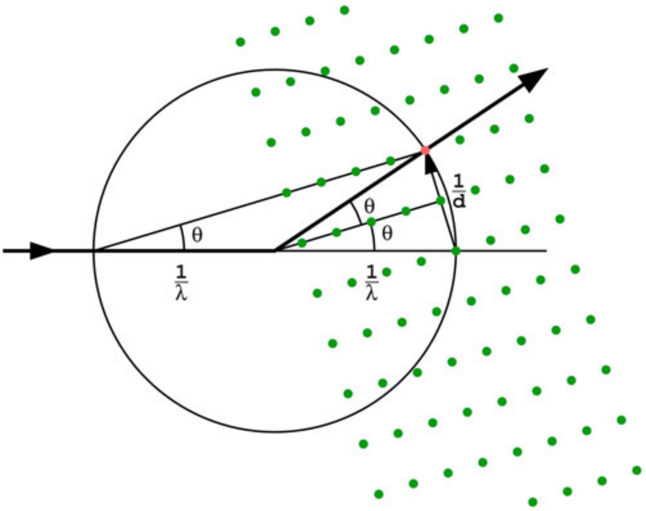


Fig. 1 The Ewald construction illustrates Bragg's law in three dimensions. The diffraction occurs at a diffraction angle, 2Θ, if the reflection, with resolution d represented by the reciprocal lattice point at a distance 1/*d* from the origin of the reciprocal space, lies at the surface of the Ewald sphere of radius 1/λ, centered at the direct X-ray beam. The graph represents the central cross-section through the three-dimensional Ewald sphere. This, and most of the other figures, are reproduced with permission from the International Union of Crystallography from Acta Crystallogr D [24]

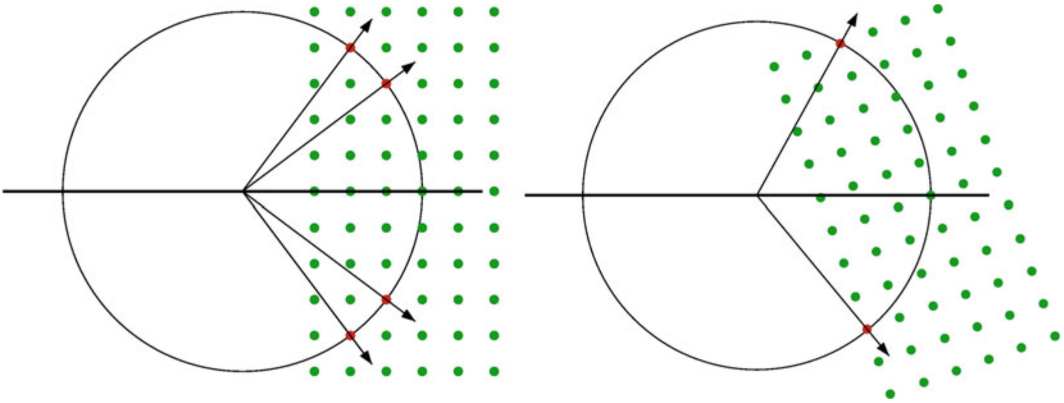


Fig. 2 To bring consecutive reflections to diffraction, the crystal, represented here as a reciprocal lattice, has to be rotated

2.1 Total Rotation Range

The total rotation range needs to be large enough to measure all reflections within the asymmetric unit of the reciprocal lattice, producing a complete data set. A range of 180° will always ensure the full completeness of the native data for all crystal symmetries, but, in many cases, completeness may be achieved earlier, avoiding unnecessary radiation damage. It is therefore beneficial to select a starting crystal orientation that ensures the full completeness in the smallest rotation range. If the crystal survives, the range may be

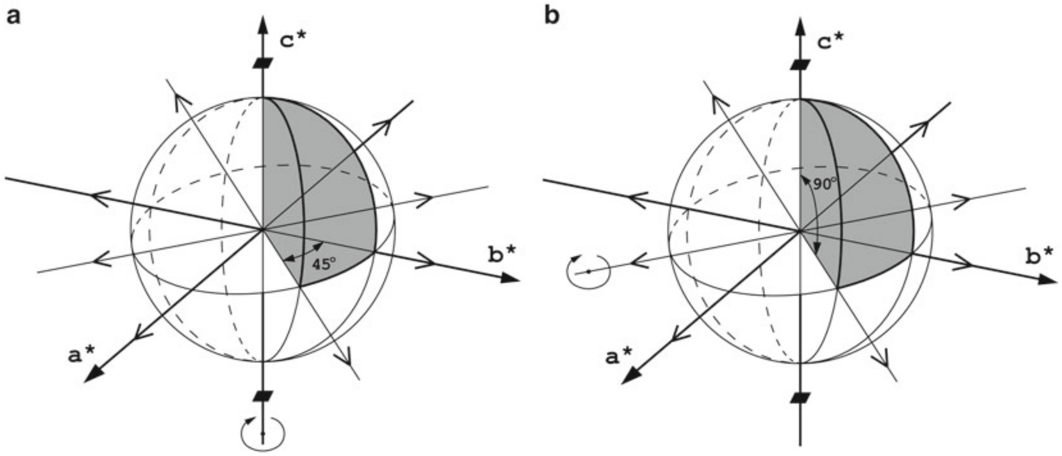


Fig. 3 The reciprocal space asymmetric unit for the crystal class 422 is shown as a shaded region. If the crystal is rotated around its fourfold axis (**a**), the width of the asymmetric wedge is 45°, but if it is rotated around any vector in the *a*, *b* plane (**b**), 90° of rotation is necessary for total completeness

extended so that more images are collected. This will increase data multiplicity, which results in increased accuracy. Selection of the optimal start and range of rotation in such a "minimalist" approach depends on the crystal symmetry and orientation. For example, when rotating a crystal of symmetry P422 around the fourfold axis, it is enough to cover 45°, but if it is rotated around a vector lying in the *a*, *b*-plane, 90° of rotation is necessary to achieve completeness of the data (Fig. 3). Moreover, if the starting orientation is incorrect, the required range may be unnecessarily extended (Fig. 4).

2.2 Mosaicity and Beam Divergence

Real crystals are built from small, mosaic blocks, slightly misoriented with respect to each other. As a result, each block diffracts in an orientation that is slightly different from that of the surrounding blocks, so that, in effect, the diffraction of a single reflection from a mosaic crystal is not instantaneous, but occurs in a small angular range, η, during crystal rotation. As a result, the direction of the diffracted beam also spans a small angular range, producing a reflection profile at the detector that is slightly widened in the angular direction. In extreme cases, if the orientation of small crystallites is completely random, as in a powder sample, each reflection forms a ring at the detector window, which is typical in the powder diffraction technique. However, mosaicity does not increase the radial width of reflections, since the Bragg angle depends only on the crystal cell dimensions.

Other effects influencing the direction of diffracted rays are X-ray beam divergence and monochromatization. The beam divergence, δ, depends on its collimation and the source size, and the spectral bandwidth $\Delta\lambda/\lambda$ of the beam depends on the properties of

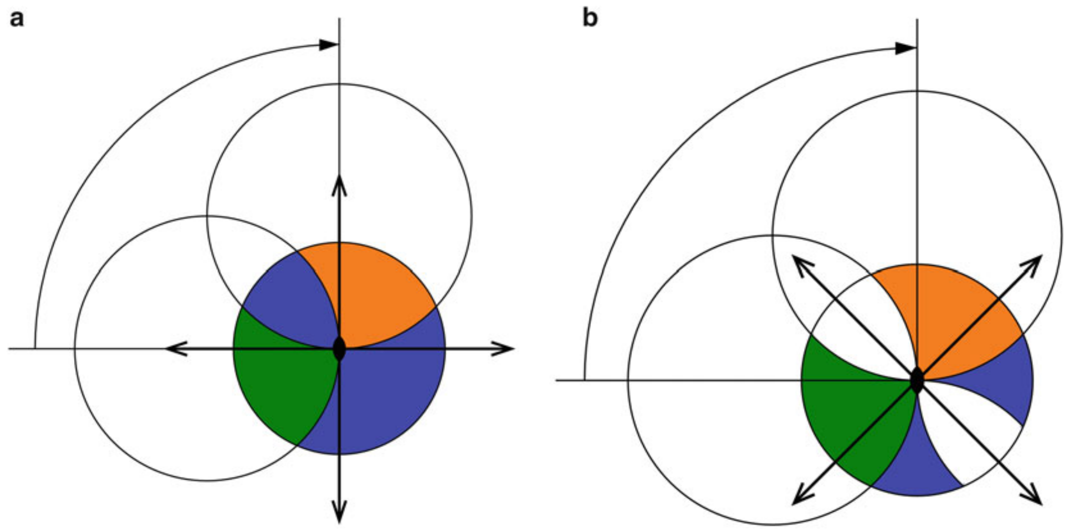


Fig. 4 The orthorhombic crystal rotated around one of its twofold axes, where reflections in the region marked in *green* are recorded on the lower half of the detector, and those in the *brown* region in the upper half of the detector window. The graphs show the central cross-section through the Ewald sphere, viewed along the spindle axis. 90° of rotation covers the full asymmetric unit, if the rotation started with the other two symmetry axes either parallel or perpendicular to the beam (**a**). In a diagonal starting orientation (**b**), the covered regions correspond to two 45°-wide symmetry-equivalent wedges, missing about 30 % of unique data

the monochromator. The primary beam is therefore not ideally parallel, and its wavelength band pass encompasses a small range. This not only increases the angular width of the reflection profiles, but also extends their radial width. The size of reflection profiles at the detector window is therefore usually larger than the size of the primary beam and dimensions of the crystal.

These effects are schematically illustrated in Fig. 5a, and their interpretation in the reciprocal space is shown in Fig. 5b. The intensity profile of the individual reflection obtained while the crystal rotates is called the rocking curve, and its width, $\Delta\theta$, depends on both crystal mosaicity and beam divergence. It also depends on the angle at which the reflection crosses the surface of the Ewald sphere, which determines how long the reflection diffracts. Additionally, the observed reflection profiles are influenced by the detector point-spread function.

2.3 Lunes and Rotation Range per Exposure

The cell dimensions of macromolecular crystals are usually much larger than the wavelength of X-rays used for diffraction experiments, and the Ewald sphere is therefore rather densely populated by the reciprocal lattice points. When a stationary crystal of protein is exposed to X-rays, a significant number of reflections will be lying at the surface of the Ewald sphere. Since the reciprocal lattice consists of planes that cross the Ewald sphere at a circle, diffracted X-rays from reciprocal points in each of these planes give

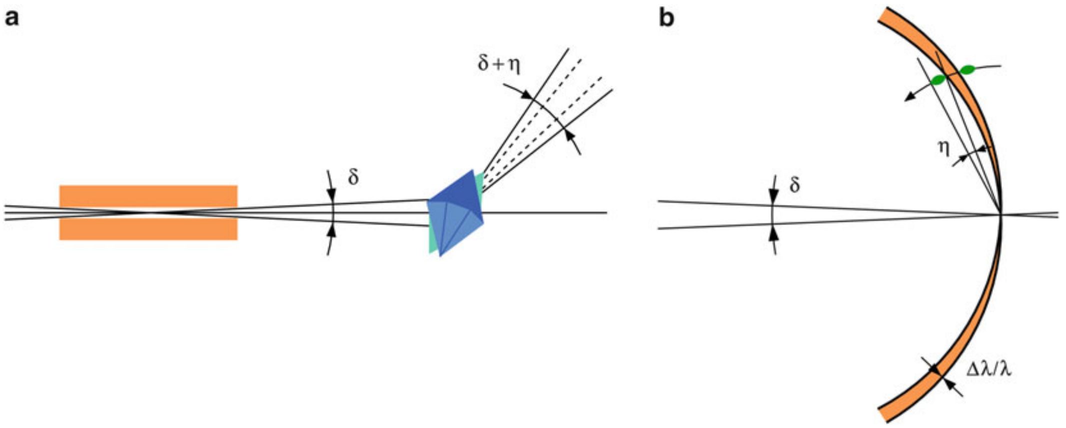


Fig. 5 Schematic illustration (**a**) of the beam divergence δ and crystal mosaicity η. The total width of the rocking curve corresponds to the sum of these two contributions. In the reciprocal space (**b**), the beam divergence is represented by a slightly rotated Ewald sphere, and the crystal mosaicity by the finite, non-zero angular size of the reciprocal lattice point. The radial size of a reciprocal lattice point depends on the uniformity of the crystal cell dimensions. The wavelength band pass, Δλ/λ, may be represented by the thickness of the Ewald sphere surface

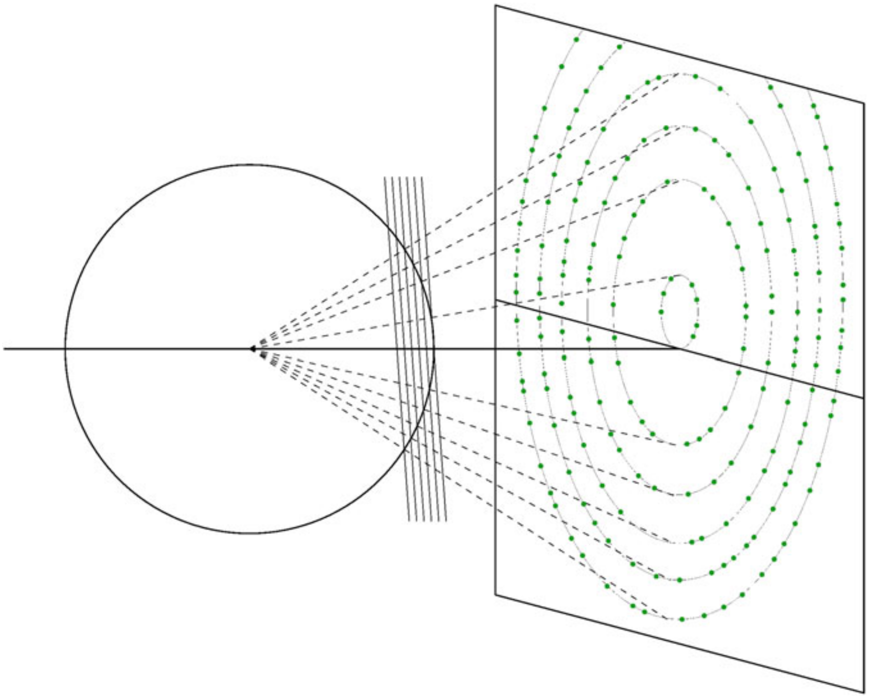


Fig. 6 If the crystal does not move during exposure, only reflections positioned at the surface of the Ewald sphere diffract. Reflections from successive parallel planes in the reciprocal lattice form a set of ellipses at the detector window, since their rays form cones having a common axis

rise to a cone of diffracted rays that produce reflections located on the flat detector window at ellipses (Fig. 6).

If the crystal rotates during exposure, all reflections between two limiting positions of each such ellipse will be recorded at the

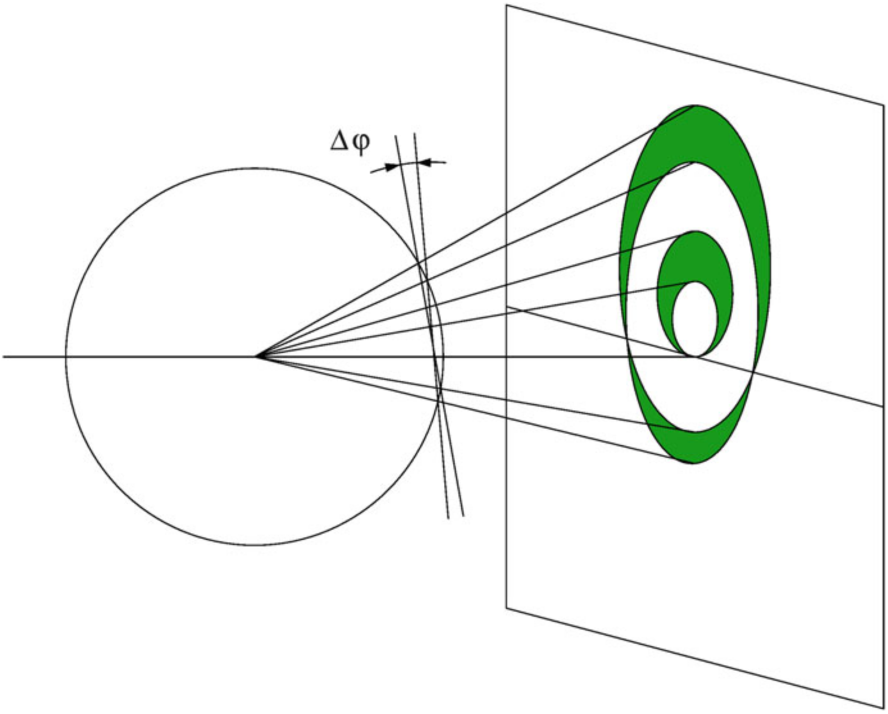


Fig. 7 If the crystal rotates during exposure, each ellipse moves accordingly, so that reflections from individual reciprocal lattice planes are grouped in lunes. The width of each lune in the direction perpendicular to the spindle axis is proportional to the amount of the crystal rotation

detector, forming a lune containing reflections from the same reciprocal lattice plane (Fig. 7). The width of each lune in the direction perpendicular to the spindle axis is proportional to $\Delta\varphi$, the angular width of the exposure. If this width is too large, the consecutive lunes will overlap at the edges of the detector, corresponding to high-angle, high-resolution reflections (Fig. 8). This situation should be avoided because the individual reflection profiles may also overlap, making the proper intensity integration impossible. The gap between two consecutive lunes depends on the distance between the reciprocal lattice planes oriented approximately perpendicular to the primary X-ray beam, which is related to the crystal unit cell dimension in this direction. The maximum allowable rotation width per image can be estimated from the formula $\Delta\varphi_{\max} = (180d)/(\pi a) - \eta$ (Fig. 9), in which d is the resolution, a is the primitive cell dimension along the beam, and $180/\pi$ converts the units to degrees. In addition, the crystal mosaicity, η, diminishes the permitted width of an image. Often, thin, plate-like crystals have their longest cell dimension in the direction perpendicular to the crystal plane and, unfortunately, such crystals tend to sit flat in the loops. It may be beneficial to use bent loops or employ the kappa goniostat and reorient the crystal to prevent the

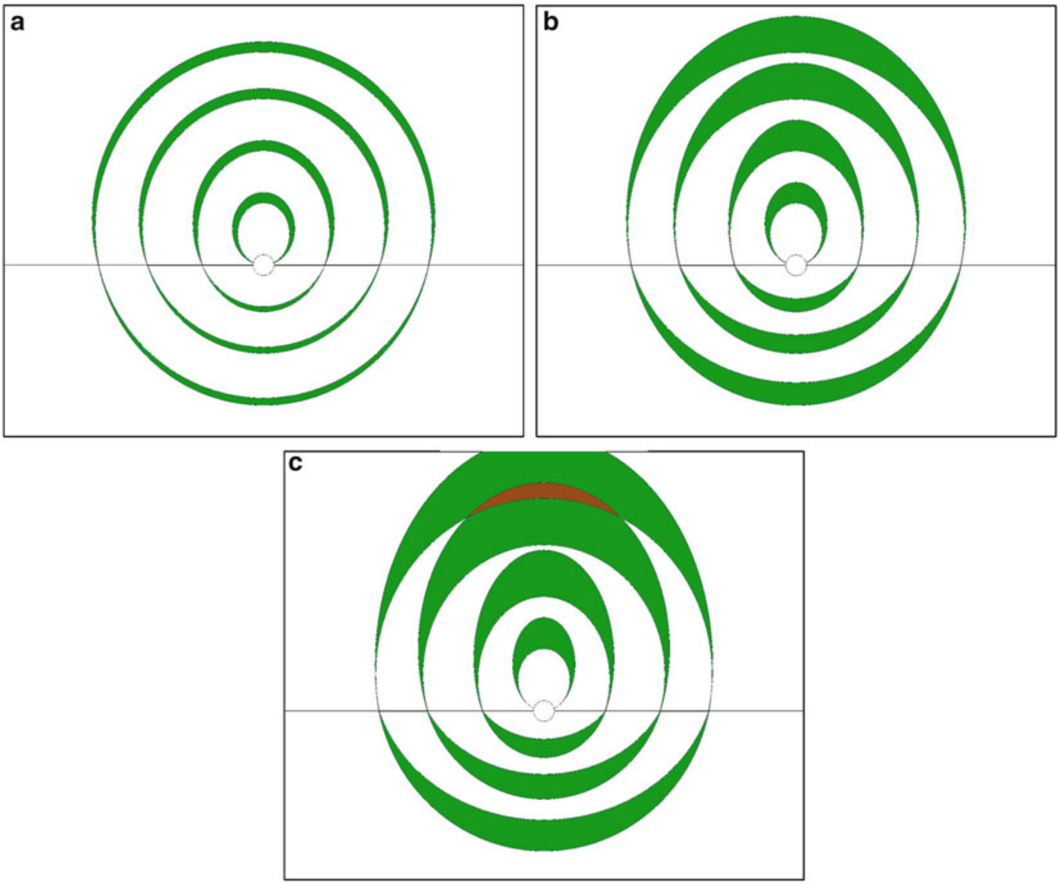


Fig. 8 The gaps between consecutive lunes depend on the distance between the reciprocal lattice planes of the same family. If the rotation range increases (**a**, **b**), the width of each lune widens, and eventually they will start overlapping at the highest diffraction angles (**c**)

longest cell axis from adopting an orientation parallel to the beam, which would otherwise cause significant overlap of reflection profiles (Fig. 10).

2.4 Fully Recorded and Partial Reflections: Wide and Fine Slicing

As mentioned previously, the diffraction by a single reflection is not instantaneous. Instead, it occurs during rotation of the crystal over a small span of time, while the reciprocal lattice point crosses the surface of the Ewald sphere. Since the data are recorded in a series of exposures corresponding to a series of consecutive narrow rotation ranges, the intensity of some reflections is spread over two or more images. Those reflections that started diffracting on one image and still diffract on the next one are called partially recorded, or simply, "partials." In contrast, those reflections whose rocking width, and therefore total intensity, are within one diffraction image are called fully recorded, or "fullys."

If the amount of rotation per image, $\Delta\varphi$, is smaller than the width of the rocking curve, $\Delta\theta$, all reflections are partials since

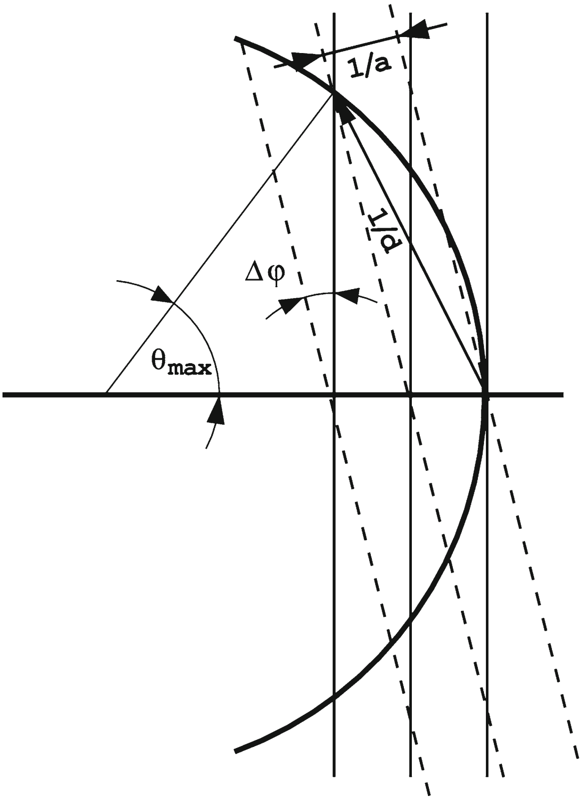


Fig. 9 To avoid overlap of the successive lunes at highest angles, the image width, $\Delta\varphi$, should be smaller than $180d/\pi a - \eta$, where d is the maximum resolution, a is the cell dimension along the beam direction, and η is the crystal mosaicity

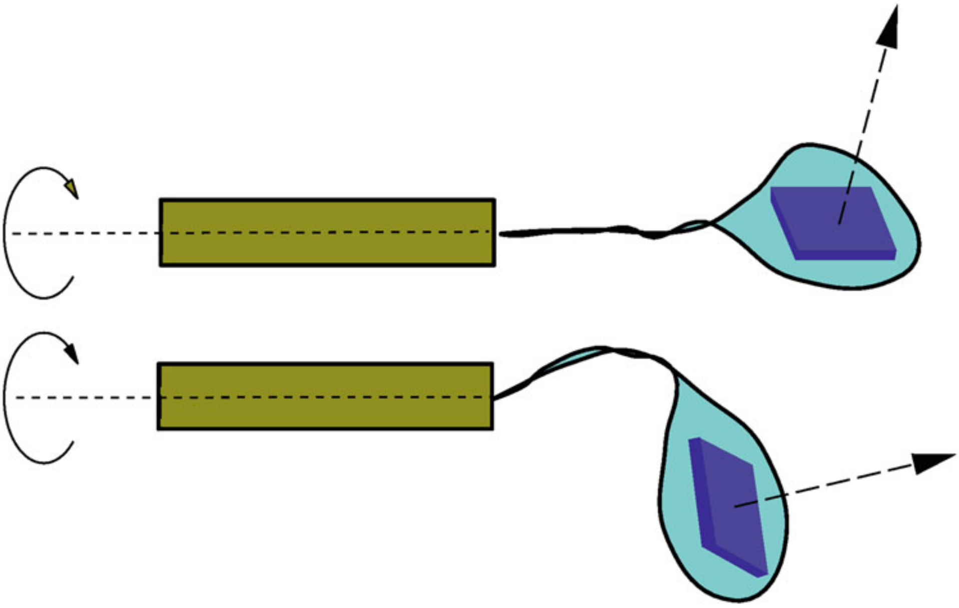


Fig. 10 For thin crystals having their longest axis perpendicular to the plate, it may be advisable to use bent loops or reorientation of the kappa goniostat because a long axis that is (approximately) parallel to the spindle will never adopt an orientation parallel to the beam. Such an orientation would lead to significant overlap of reflection profiles

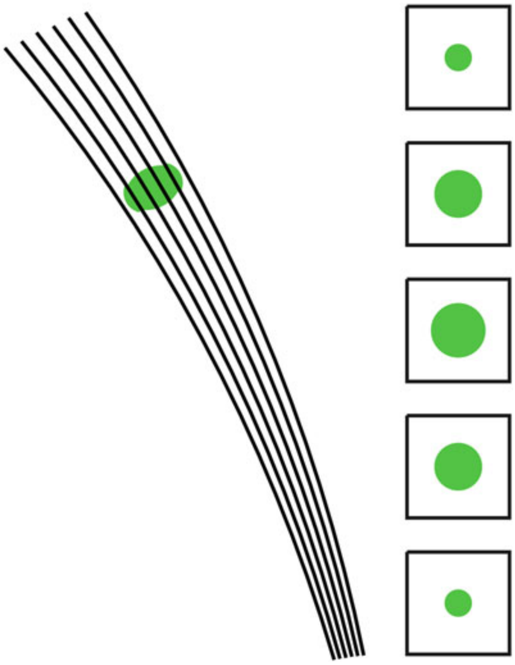


Fig. 11 If the rotation range is much smaller than the crystal mosaicity, each reflection is spread over several images, and it is possible to build a three-dimensional profile in the so-called "shoe-box"

each reflection is wider than the angular width of the exposure. This leads to two different approaches for data collection, wide slicing, when $\Delta\varphi \approx \Delta\theta$, and fine slicing, when $\Delta\varphi \ll \Delta\theta$. These two methods utilize different ways of integrating reflection intensities.

In the wide-slicing technique, the intensity is integrated in all detector pixels of the individual reflection profile, and the background level is estimated from surrounding pixels within each recorded diffraction image. For partial reflections, the intensities estimated from subsequent images are simply added. In the fine-slicing approach, it is possible to build the reflection profile and to estimate background in three dimensions, i.e., the detector x and y coordinates and the "perpendicular" direction of the spindle φ rotation, using appropriate pixels from consecutive images (Fig. 11) [8].

If the image width, $\Delta\varphi$, is significantly larger than the reflections width, $\Delta\theta$, the reflection intensities are recorded at only a fraction, $\Delta\theta/\Delta\varphi$, of the total exposure time, while the background accumulates during the whole exposure. As a consequence, the resulting signal-to-noise ratio in wide slicing is worse than in the fine-slicing mode.

In the wide-slicing mode, each lune contains fully and partially recorded reflections. Fullys are located in the middle of each lune, and partials are close to the edges. Partials at the upper edge of each lune appear at the next exposure, at the lower edge of the

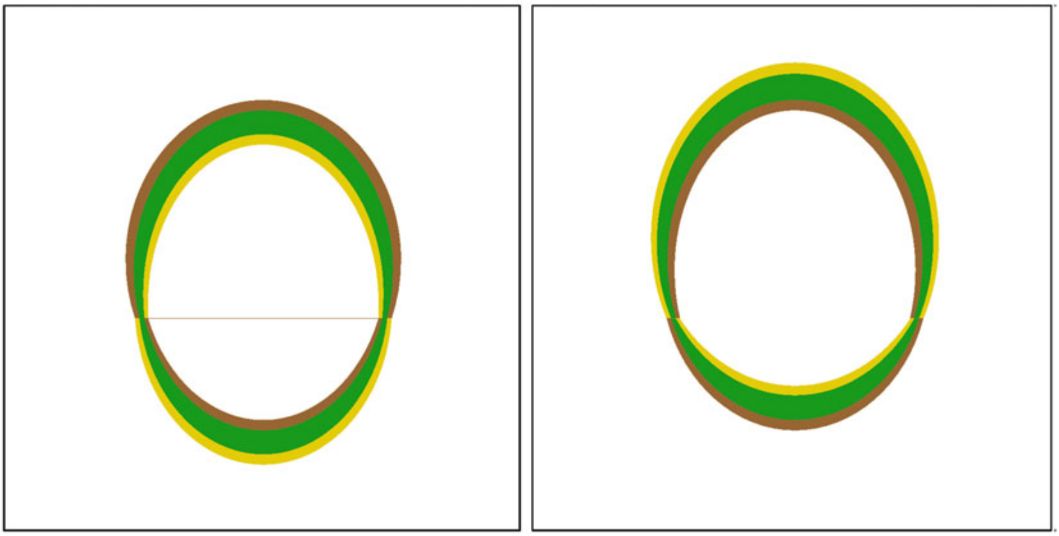


Fig. 12 Fully recorded reflections occur in the middle of each lune. Partials remaining from the previous image are located near the lower edge, and those that just started diffracting occur near the upper edge of each lune, provided that a crystal rotates upwards at the side closer to the detector

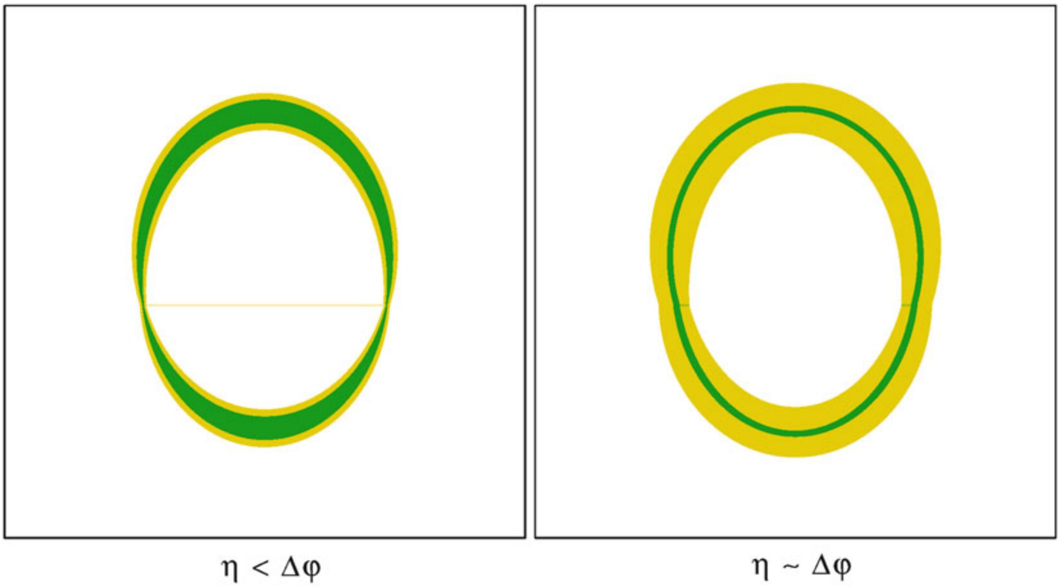


Fig. 13 High mosaicity increases the number of partials, making each lune wider. Characteristically, with low mosaicity, the edges of each lune are sharply defined; with high mosaicity, the reflection intensities gradually fade away without making well-defined lune edges

corresponding lunes (Fig. 12). The appearance of the lune's edges depends on the crystal mosaicity. If each lune has well-defined, sharp edges, the mosaicity is small, but reflection intensities that fade away gradually without forming well-defined lune edges suggest that the mosaicity is high (Fig. 13).

The fine slicing may lead to more accurate intensity estimations, but at the cost of exposing many more images. This factor is important if the detector readout time is relatively large. Collecting fine sliced data at a synchrotron beamline with an Image Plate scanner with 1-s exposures and 30 s of detector cycle time is obviously not economical. In contrast, with a pixel array detector with readout time in the milliseconds range, the fine-slicing mode should be the method of choice.

2.5 Blind Region

Even if the total rotation reaches 360°, some reciprocal lattice points lying close to the rotation axis will have no chance to cross the Ewald sphere (Fig. 14). Reflections in this "blind region," or "cusp", cannot be measured in one rotation pass of data collection. The width of the blind region depends on the curvature of the Ewald sphere and therefore on the X-ray wavelength (Fig 15). Using a short wavelength minimizes the fraction of reflections lost in the blind region. For a wavelength of 1 Å, reflections lost do not exceed about 2 % at 2.0 Å resolution and about 8 % at 1.0 Å resolution. In fact, if the unique crystal axis is mis-set from the direction of the spindle axis, all reflections within the blind region will have their symmetry mates in a different region of reciprocal space, and the overall data completeness will not suffer (Fig. 16). The blind region negatively affects the data completeness only if the crystal is oriented and rotated around its unique symmetry axis or if it has *P*1 symmetry.

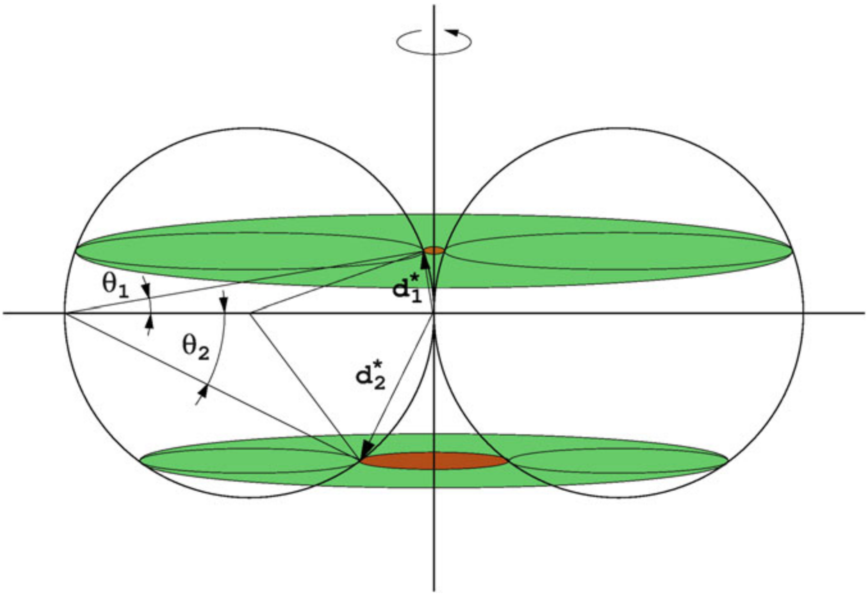


Fig. 14 Even after 360° rotation, some reflections in the blind region (shown in *brown*), close to the rotation axis, will never cross the surface of the Ewald sphere. The blind region is narrow at low resolution, but is significant at high resolution, numerically comparable to the X-ray wavelength. The fraction of reflections lost in the blind region at diffraction angle θ is 1 – cosθ. (Color figure online)

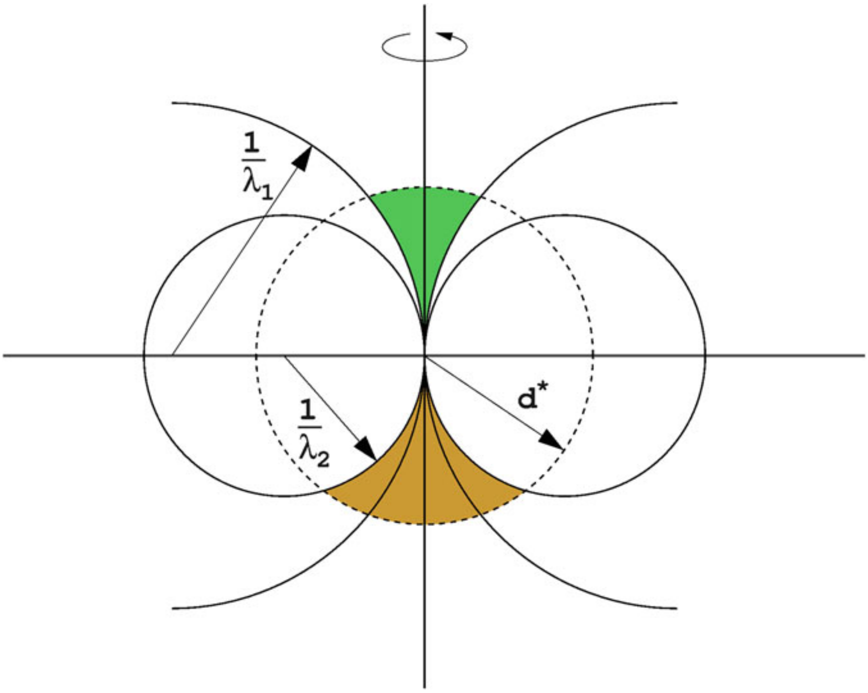


Fig. 15 The width of the blind region depends on the wavelength, defining the curvature of the Ewald sphere. At a short wavelength, the blind region is smaller than at a long wavelength

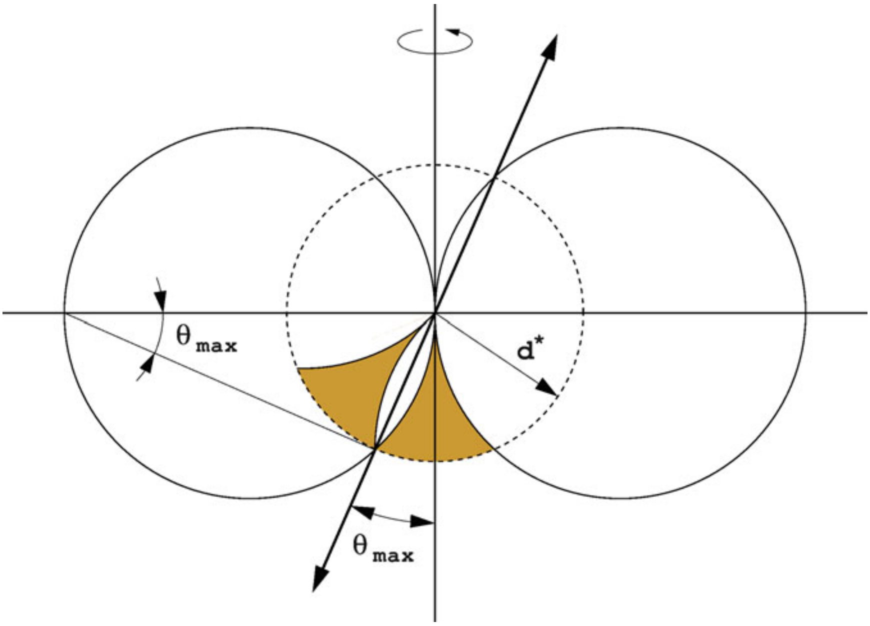


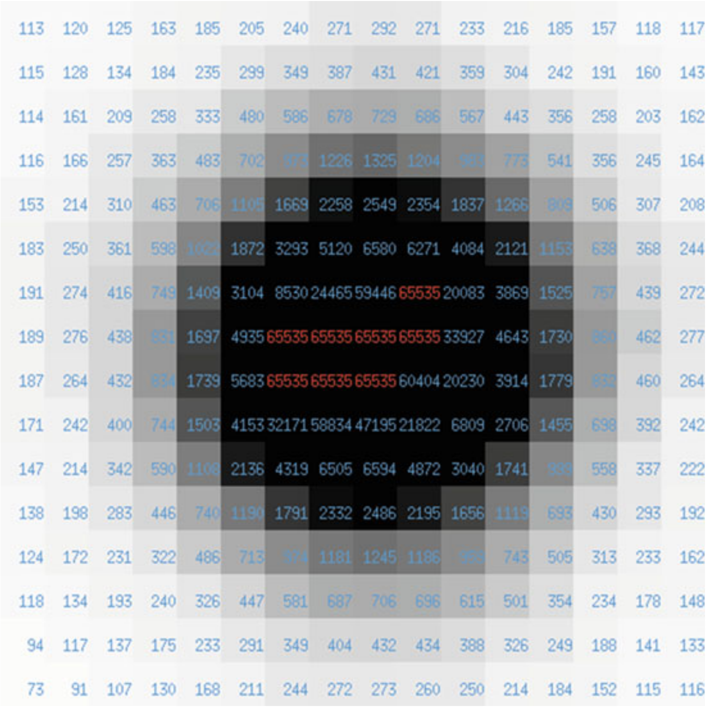
Fig. 16 If the crystal has one unique symmetry axis, it is beneficial to mis-set it from the direction of the spindle axis by at least θ_{max}. By doing so, all reflections in the blind region will have their symmetry mates in the measurable region of the reciprocal space

2.6 Overloaded Detector Pixels

Each detector has a certain limit of the dynamic range, i.e., maximum intensity that can be measured and stored in a single pixel. For example, most of CCD detectors store numbers as 16-bit integers, so that the maximum pixel value is $2^{16}-1=65{,}535$, and higher intensities are truncated to this value (Fig. 17). The PILATUS detectors work with 20-bit numbers, and their numerical dynamic range is about 1 million.

Well diffracting crystals require a high X-ray dose or sufficiently long exposures to adequately measure all high-resolution reflections, often resulting in a number of strong, low-resolution reflections having overloaded pixels in their diffraction profiles. These strongest reflections play an important role in the anomalous and molecular replacement phasing procedures. They should be adequately measured in a separate rotation pass of data collection, with shorter exposures or attenuated beam intensity. Such a "low-resolution" pass may cover only low-resolution data, using longer detector distance and wider rotation per image. Reflections from all passes should then be scaled together.

If the initial exposure suggests that an additional, low-resolution pass may be required, it is better to perform it before the high-resolution pass. It is beneficial to measure the most important reflections while the crystal is not significantly affected by radiation



113	120	125	163	185	205	240	271	292	271	233	216	185	157	118	117
115	128	134	184	235	299	349	387	431	421	359	304	242	191	160	143
114	161	209	258	333	480	586	678	729	686	567	443	356	258	203	162
116	166	257	363	483	702	873	1226	1325	1204	983	773	541	356	245	164
153	214	310	463	706	1105	1669	2258	2549	2354	1837	1266	809	506	307	208
183	250	361	598	1022	1872	3293	5120	6580	6271	4084	2121	1153	638	368	244
191	274	416	749	1409	3104	8530	24465	59446	65535	20083	3869	1525	757	439	272
189	276	438	831	1697	4935	65535	65535	65535	65535	33927	4643	1730	860	462	277
187	264	432	834	1739	5683	65535	65535	65535	60404	20230	3914	1779	832	460	264
171	242	400	744	1503	4153	32171	58834	47195	21822	6809	2706	1455	698	392	242
147	214	342	590	1108	2136	4319	6505	6594	4872	3040	1741	939	558	337	222
138	198	283	446	740	1190	1791	2332	2486	2195	1656	1119	693	430	293	192
124	172	231	322	486	713	974	1181	1245	1186	958	743	505	313	233	162
118	134	193	240	326	447	581	687	706	696	615	501	354	234	178	148
94	117	137	175	233	291	349	404	432	434	388	326	249	188	141	133
73	91	107	130	168	211	244	272	273	260	250	214	184	152	115	116

Fig. 17 A detailed profile of an overloaded reflection, with several central pixels having the maximum tolerated value 65,535

damage, inevitably incurred when exposures are longer. The low-resolution pass involves much less damage, and the subsequent high-exposure pass is therefore not significantly influenced.

In this context, using the fine-slicing mode is also beneficial because the strongest intensities are split between several images, and the probability of overloads is considerably diminished.

2.7 Alternative Indexing and Twinning

In certain crystal classes, reflections can be indexed according to more than one permitted, but not equivalent, schemes. This occurs when the symmetry of the crystal class (point group) is lower than the symmetry of the lattice, as, e.g., for crystals with polar axes that can be directed in two ways (Fig. 18). The affected point groups are 4, 3, 321, 312, 6, and cubic 23, including space groups with all combinations with screw axes. For a single pass of data collection, it is immaterial how all the reflections are indexed, but for merging or comparing data from multiple passes or separate crystals (including derivatives), it is important to preserve the same indexing scheme in all contributing sets of data. This effect may also occur in other symmetries, if some cell parameters serendipitously adopt certain particular values and the crystal lattice "pretends" to show higher symmetry than that of the crystal structure.

If, at the stage of data collection, the crystal structure and true symmetry are unknown, then in cases of tetragonal, trigonal, hexagonal and cubic lattices it is always safer to assume that the crystal has symmetry lower than holohedry (4 instead of 422, 3 instead of 622, and 23 instead of 432) and adjust the data collection strategy accordingly.

It is advisable to always start the data collection at the appropriate optimal crystal orientation to achieve completeness after

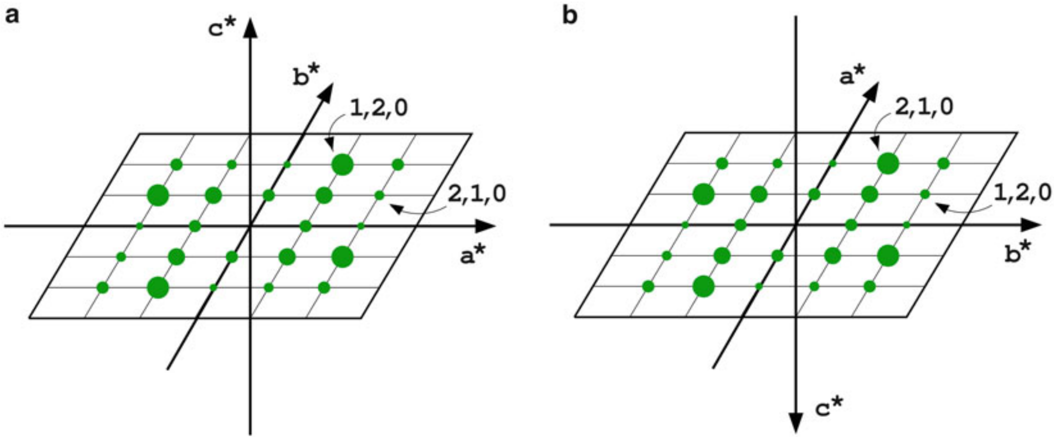


Fig. 18 In polar space groups, as, for example, in P4, reflections can be indexed in two permitted, but not equivalent, ways, with the unique polar axis directed in one (**a**) or the opposite (**b**) way. This has to be taken into account in merging reflections from different data sets

minimal total rotation range, but to continue collecting 180° of data. If completeness is achieved earlier, the images affected by radiation damage can be discarded, but this approach may be beneficial if the crystal symmetry is found to be lower than suggested by the initial indexing.

The same crystal classes can be affected by merohedral twinning [9], which occurs when the individual crystalline specimen contains separate, alternatively oriented domains, and non-equivalent reflections from both domains overlap precisely. In such cases, the same considerations apply, since the real crystal symmetry is then lower than apparent from indexing and even from scaling the data. In perfect merohedral twins, when the twinning fraction is 1/2, i.e., the irradiated volume of both twin domains is equal, the scaling statistics (R_{merge}) may suggest a high degree of symmetry, whereas in reality, the crystal has lower symmetry. Data from twinned crystals are characterized by an intensity distribution that is different from the "normal" Wilson statistics, with smaller fractions of very weak and very strong reflections, which are apparent from, e.g., N(z) or H tests.

It is always advisable to test diffraction data for twinning early on, because more than the expected rotation range may be necessary. It is possible to quickly test data for twinning at the dedicated "Merohedral Crystal Twinning Server" http://nihserver.mbi.ucla.edu/Twinning/.

2.8 Radiation Damage

Radiation damage incurred to protein crystals, especially at the bright contemporary synchrotron beamlines, results in significant degradation of diffraction data quality. Even at cryo-temperatures of about 100° K after absorbing about 20–43 MGy, the total intensity of all diffracted reflections diminishes to half of the original value [10, 11]. The first to suffer are the highest-resolution reflections, and 1 MGy of the absorbed dose increases the data-scaling B factor by about 1 Å^2, but the intensities of the low-resolution, strongest reflections also change as a result of structural rearrangements and chemical modifications (breakage of disulfide bridges, decarboxylation of acids, etc.). The effects of radiation damage, therefore, degrade not only the data resolution and quality, but may also be responsible for potential misinterpretation of certain, fine structural features, such as partially occupied ligands, or of the behavior of functionally important residues.

Cryo-cooling and, in certain cases, the use of radical scavengers diminish the secondary effects resulting from the diffusion of certain active species throughout the crystal. However, primary radiation damage, i.e., the immediate effect following the absorption of X-ray quanta, as a physical phenomenon is inevitable [12]. In practice, limiting the radiation damage can be achieved only by reducing the exposure time or attenuating the intensity of the X-ray beam. As pointed out previously, a certain degree of damage is allowable if the data are to be used for final model refinement, but for anomalous phasing applications, it should be avoided.

It is therefore advisable to check the data for radiation damage early in the data collection process. Contemporary data processing programs allow integration and merging of collected data almost in parallel to the image acquisition. The existing strategy programs, e.g., BEST [5], RADDOSE [11], are able to suggest the appropriate level of exposure, allowing the collection of complete data within the selected total absorbed dose.

The useful criteria for radiation damage are the scaling B factors and R_{merge} values. Often the degradation of the reflection profiles and the loss of high resolution are apparent by visual inspection of the diffraction images. The R_{merge} and χ^2 values of the individual images may reveal characteristic behavior, with the highest values at the beginning and end of the range, and the smaller values in the middle (Fig.19). This distribution of values occurs because average intensities are most similar to those recorded in the middle of the session and therefore most distant from those measured at the start and end of the session.

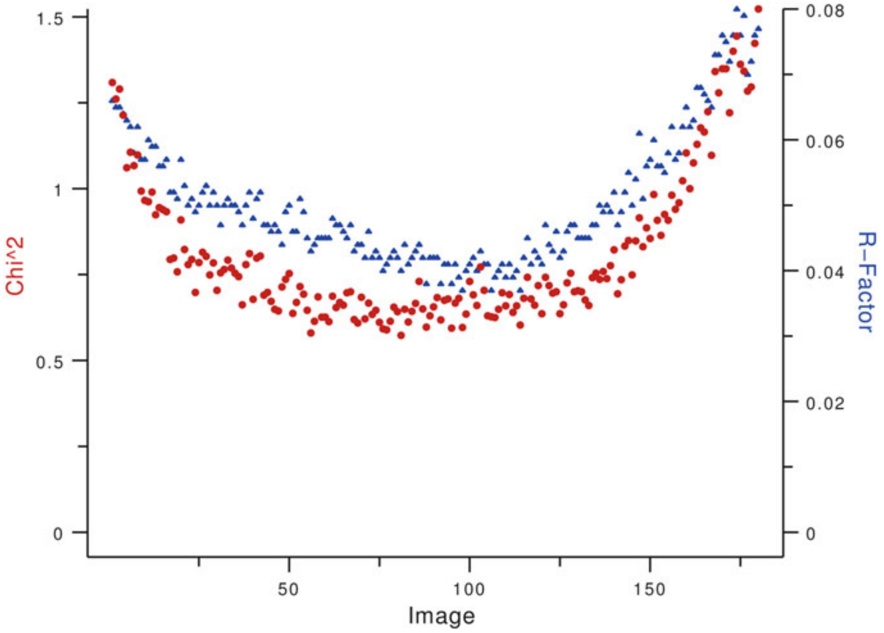


Fig. 19 R_{merge} and χ^2 values for individual images from a severely radiation-damaged crystal. Both of these values are large for images at the beginning and end of the whole set, when the intensities are most different from values averaged over the entire set. In contrast, the intensities from images recorded in the middle, have intensities similar to the average values for the entire set

3 Practical Protocols

3.1 Most Common Approaches

Two approaches are most often executed in practice, especially within structural genomics projects. The first applies to proteins with no known similar structure in the PDB that could serve as an MR search model; the second is used for cases in which a suitable search model is available. Often the crystallized protein has selenomethionine introduced to its sequence, even if it can be expected to be solved by MR. The advisable protocols for these two cases are similar, with somewhat different priorities in the specific details.

Obviously, it is necessary to start the data collection session by executing the necessary preliminaries, such as selecting the appropriate wavelength, accurately centering the crystal (and beam) at the rotation spindle axis, etc. At the beginning, a couple of orthogonal (separated by 90°) images should be exposed with typical conditions (e.g., 0.5° rotation, detector distance set for 2 Å resolution, and modest beam intensity) and carefully inspected visually. Often, such inspection allows the crystal to be discarded immediately if the observed images cannot be interpreted as forming a single, or at least a clearly identifiable, lattice of reflections. The initial images should be indexed and (assuming it was successful) further decisions about strategy and data collection parameters adjusted, preferably with the use of a strategy program, on the basis of the initial interpretation of these images, i.e., the estimated crystal symmetry, cell dimensions, mosaicity, orientation, resolution limit, etc. As emphasized above, it is advisable to avoid overexposing the crystal; in practice, there should be no more than a few overloaded detector pixels present in each image. If necessary, for very-well-diffracting crystals, the low-resolution pass should be executed first, before the second pass, to encompass all weak, high-angle reflections. The results of the low-resolution pass may also suggest improved parameters for the second pass.

Attention should be paid to the optimal selection of the start range of rotation, so that data completeness may be achieved after minimal total rotation. However, it is beneficial to continue recording images, which, if the crystal is not excessively damaged by the radiation, are useful for the enhancement of data multiplicity and possibly for identifying cases of pseudosymmetry or twinning.

It is strongly advised to proceed with the integration and initial merging of recorded intensities immediately after the start of data collection. In fact, there is no excuse for not doing so. Any initial errors and misinterpretations can then be adequately and rapidly corrected. Moreover, the structure solution using the SAD or MR approaches can be attempted rapidly, when the crystal is still on the goniostat, and decisions can be made about collecting additional data at the same or a different wavelength (moving from SAD to MAD). If the results obtained from one crystal are not satisfactory, another specimen can be utilized immediately. Automatic programs such as xia2 [13], EDNA [14], etc. can be useful for achieving better throughput; however, one should carefully inspect the characteristics and quality of the data in the data processing logfiles.

Different procedures must be applied when many sets of images are recorded “blindly” from a number of similar crystals with the intention of interpreting them later. In such an approach, it is only possible to evaluate and select the best set of images and process them without the possibility of “run time” intervention and feedback. This kind of approach considerably limits the human effort at the expense of using a large amount of beam time. It is more beneficial to treat the diffraction data collection as a scientific process, not as a mere technicality.

3.2 Choice of Wavelength

Data intended for molecular replacement can be measured with any wavelength. Most of the synchrotron beamlines perform optimally at about 1 Å, and this wavelength is appropriate for MR applications and for collecting the ultimate data for structure refinement. The wavelength may need to be shorter only if the crystals diffract to atomic resolution; otherwise, the limitations of the shortest available detector distance may preclude achieving high enough diffraction angles.

Collecting the SAD data requires that the wavelength is in the region providing significant anomalous signal from the anomalous scatterers present in the sample. The maximum f'' value corresponds to the peak point of the fluorescence spectrum, with the caveat that not all elements have their absorption edges in the wavelength region available at most of the macromolecular synchrotron beamlines (Fig. 20). One can either set the wavelength to the peak value suggested from the spectrum, or use a wavelength in the high-energy, remote region, 50–100 eV above the expected edge value, without recording the spectrum. For utilizing lighter anomalous scatterers which have their edges beyond the wavelength region available at synchrotrons (Ca, K, Cl, S, P), longer wavelengths are preferred, in the region of 1.7–2.1 Å [15] and, if available, with a helium path between the crystal and detector, to diminish absorption and scattering of X-rays by air. To obtain a good anomalous signal from sulfur in native proteins, it is advisable to extend data redundancy, but with a significantly attenuated beam.

For the MAD work, it is necessary to record the fluorescent spectrum. The accurate values of the peak and edge wavelengths, and estimations of the anomalous corrections f' and f'' can be obtained using the program CHOOCH [16]. There is no consensus among the community about the best protocol for performing the MAD experiment, except that everybody agrees that the level of exposure should not be too ambitious. One option is to collect data at three wavelengths (in any order) with modest redundancy. Another protocol includes collecting only at the edge and remote

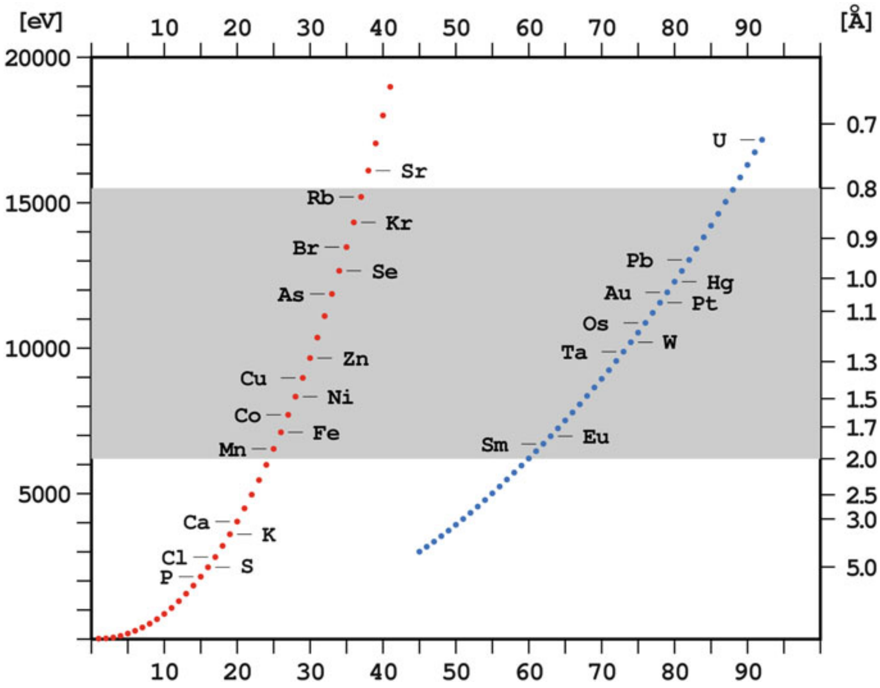


Fig. 20 K (*red*) and L_{III} (*blue*) absorption edges of various elements. *Symbols* of the most often used anomalous scatterers are shown explicitly. The X-ray wavelength (energy) region available at most synchrotron beamlines is *shaded*

wavelengths with extended redundancy, avoiding the peak wavelength altogether, where the most absorption and radiation damage would be incurred. This consideration is very important for elements characterized by L_{III} edges with extremely large f' values at peak wavelengths, such as lanthanides (Sm, Eu, Gd) and tantalum (often used as $Ta_6Br_{12}^{2-}$ cluster). On the other hand, mercury does not show any white line in its spectrum, so that only the edge and remote wavelengths are useful. The Hg-derivative data are especially susceptible to radiation damage, since the Hg bonds to cysteine break very easily.

The most common anomalous scatterer for SAD and MAD phasing is selenium, genetically introduced to proteins in the form of selenomethionine (SeMet). This procedure is standard for proteins expressed in bacteria, but it may be difficult to introduce SeMet to proteins obtained through other protocols. If the protein naturally contains metals such as Zn, Cu, Fe, or Mn, they can be used for anomalous phasing. Proteins often naturally coordinate calcium (Ca), and it may be possible to exchange Ca for lanthanides, providing a very significant anomalous signal. It is also possible to soak protein crystals in heavy-metal salts according to classic derivatization protocols or by short-soaking approaches [17]. The short soaking in cryo-solutions containing bromides (for MAD) or iodides (for SAD) provides another possibility [18]. Crystals of large structures such as multiprotein or protein:DNA complexes can be derivatized with polynuclear metal clusters, such as, for example, $[Ta_6Br_{12}]^{2-}$, $[PW_{12}O_{40}]^{3-}$, providing very strong anomalous signals from the multicenter "superatoms," especially at low resolution [19].

3.3 Quality Criteria

Several criteria are commonly used to judge data quality, but not all of them are equally useful or statistically accurate. The traditional, and obligatorily quoted, $R_{merge} = (\Sigma_{\mathrm{hkl}}\Sigma_i|I_{\mathrm{i}} - <\mathrm{I}>|)/(\Sigma_{hkl}\Sigma_{\mathrm{i}}I_{\mathrm{i}})$ is not statistically valid, since it does not take into account the effects of multiple measurements. More informative forms of agreement factors have been proposed, such as $R_{\mathrm{meas}} = (\Sigma_{\mathrm{hkl}}[n/(n-1)]\Sigma_i|I_i - <\mathrm{I}>|)/(\Sigma_{\mathrm{hkl}}\Sigma_i I_i)$ [20], and $R_{\mathrm{pim}} = (\Sigma_{\mathrm{hkl}}[1/(n-1)]\Sigma_i|I_i - <I>|)/(\Sigma_{\mathrm{hkl}}\Sigma_i I_i)$ [21]. Unfortunately, these are not universally adopted in publications or in the PDB. Another measure of data quality is the average $I/\sigma(I)$ ratio; however, it is not always easy to properly estimate reflection uncertainties, $\sigma(I)$, by using counting statistics, since two-dimensional detectors do not measure individual X-ray quanta but reproduce values proportional to their number. High multiplicity provides a better estimation of uncertainties from the real spread of individual measurements around the average value of intensity. Of course, data completeness and multiplicity also provide important information. The useful and statistically valid new criterion recently proposed is $\mathrm{CC}_{1/2}$ [22], the correlation coefficient between two halves of the data set, scaled as a whole and merged in two randomly selected parts. It is important to inspect the values of the quality criteria at the highest-resolution range. The commonly accepted data resolution limit used corresponds to $I/\sigma(I)$ of about 2.0, but there are indications that this criterion should be significantly relaxed [22].

Anomalous data are also judged by the average Bijvoet ratio $\Delta F_{\mathrm{anom}}/F$ (as a function of resolution) and $\mathrm{CC}_{\mathrm{anom}}$, the correlation coefficient between signed anomalous differences in two randomly split halves of the data. Anomalous signal useful for phasing exists in resolution ranges where $\mathrm{CC}_{\mathrm{anom}}$ is higher than 30 % [23].

3.4 Potential Problems and Their Remedies

There are several reasons for failure of data collection experiments. Obviously, unsatisfactory crystal quality is the most common scenario, but often, apparently difficult situations may be remedied after proper interpretation.

3.4.1 Incorrect Beam Center

Incorrect detector coordinates of the beam center is a common source of failure of autoindexing. The beam center is the location of the (0, 0, 0) reflection and the place at the detector where a direct beam intercepts its front window if the beam stop is removed. For successful indexing of diffraction images, accurate beam center coordinates are crucial. Deviation of the beam center by a small

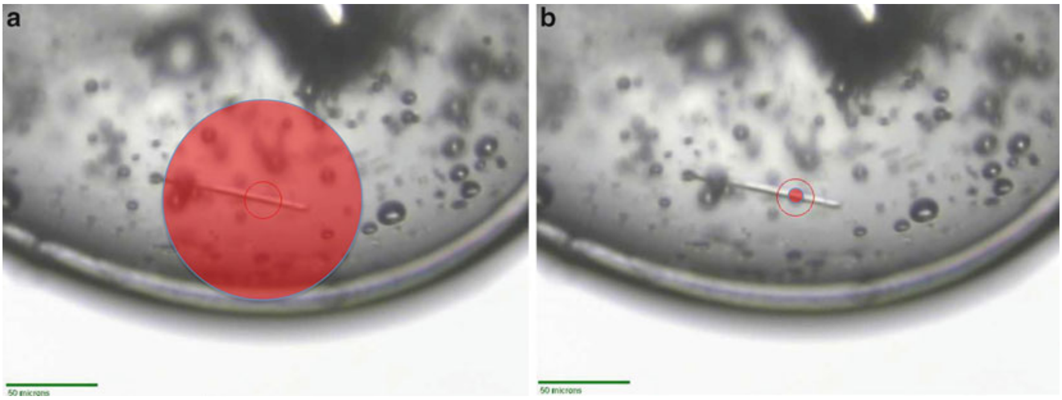


Fig. 21 (**a**) If the beam size is much larger than the crystal, there will be a significant additional scatter of the cryo-solution and loop, unnecessarily increasing the level of background at the diffraction images. (**b**) If the beam size is adjusted to the crystal size, the background level is lower, and the signal-to-noise ratio of the data is therefore higher

amount (several pixels) can be tolerated and refined by indexing programs. However, large deviations (more than half of the spot separation) result in failure in indexing. Hence, the failure rate increases with the length of the cell dimensions, where the spot separation is small. If the detector rail is not precisely parallel to the direct beam direction, the beam center at two different crystal-to-detector distances may not be the same. It is advisable to confirm, or obtain the accurate values of the beam center from the beam line staff or from previous experiments.

3.4.2 Use of Small Beam for Nonuniform Crystals

A large beam and a big crystal of good quality always provide the best data. For smaller crystals, it is advisable to adjust the beam to the crystal size (Fig. 21), to avoid unwanted background scattering from noncrystalline surroundings (excess solvent, fiber loop, air). However, this rule cannot be easily applied if the crystals are very elongated or nonuniform within their whole volume. For long, thin, needle-like crystals, it is possible to collect data at several discrete spots along the length of the needle and scale them together to generate a complete data set (Fig. 22a). A better way is to perform the so-called "vector scan" or "helical scan," available at several minibeam-capable synchrotron facilities, where each image is collected at a slightly different spot on the crystal and successively translated along the length of the crystal (Fig. 22b). This process also maximizes the use of the whole crystal volume, thereby reducing the radiation damage.

Good single crystals are capable of providing high-quality data, no matter which part of the crystal is used for data collection. However, in practice, crystals may be nonuniform in quality, non-homogenous, cracked, warped (because of the meniscus force of the cryo-solution), etc. The use of a minibeam of a few microns

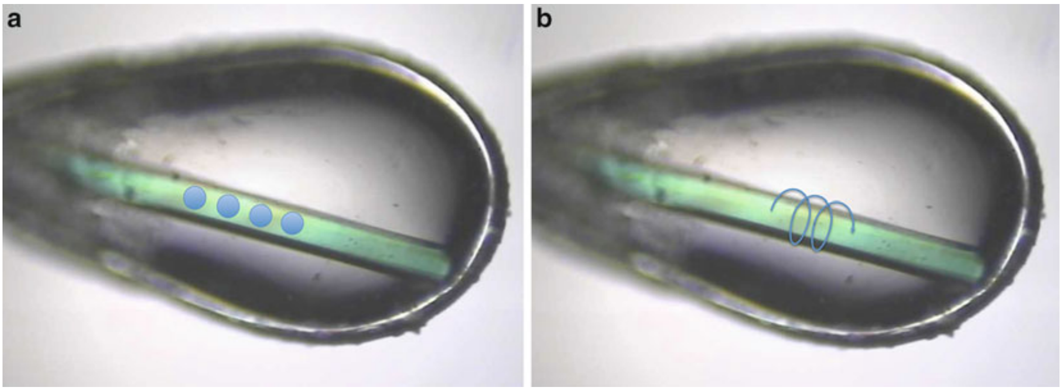


Fig. 22 With long, thin crystals it is possible increase the signal-to-noise ratio and to minimize radiation damage by acquiring data with a small beam at several points along the crystal (**a**), or with the so-called helical data collection approach (**b**), in which the crystal is slowly translated during rotation and the beam moves between the two ends of the crystal

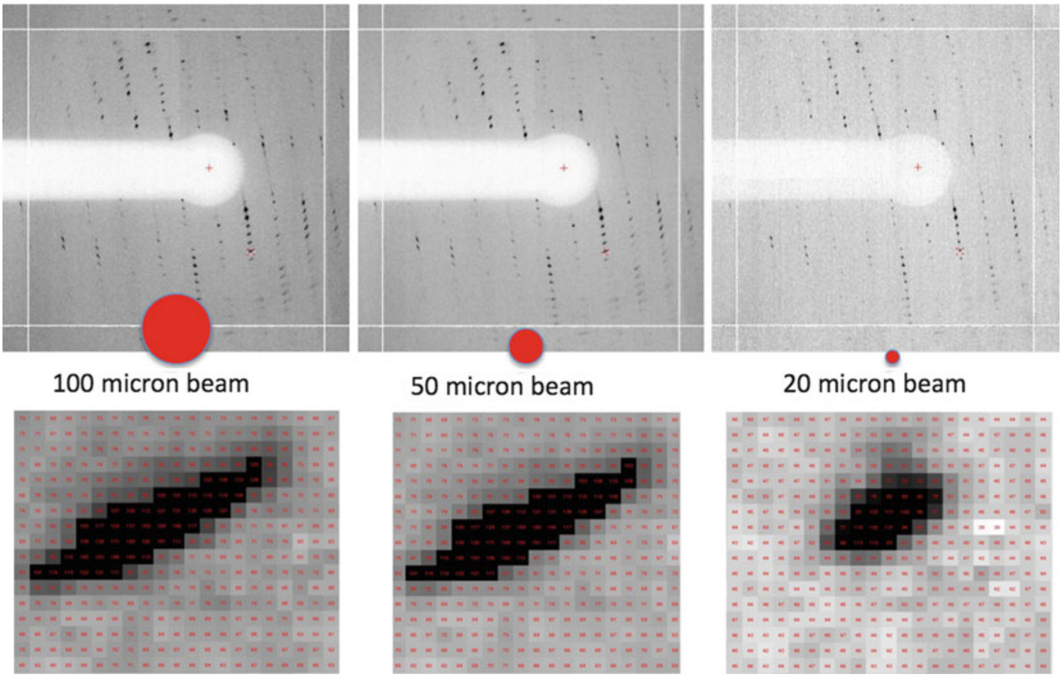


Fig. 23 Analogous diffraction images recorded from a large, warped crystal using a large (100 μm, *left),* medium (50 μm, *middle*), and small (20 μm, *right*) beam. With the larger beam, the reflection profiles are elongated and diffused, suggesting very high mosaicity. A small beam results in more uniform reflection profiles and a cleaner image background

in size can provide useful diffraction data from such samples. Using microdiffraction techniques, one can evaluate the quality of diffraction on different parts of a large crystal and collect data on the best part. Figure 23 shows diffraction spot profiles from a warped crystal.

As can be seen, with a 100-μm beam, the reflection profiles are very streaky, making it difficult to estimate their Bragg intensities. The spot profiles improve when a smaller beam is used. With a 20-μm beam, the spots are less streaky and usable for integration.

Acknowledgements

K.R.R. is supported by a grant from National Institute of General Medical Sciences (8 P41 GM103403-10) of the National Institutes of Health. Z.D. has been supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

References

1. Dauter Z, Wilson KS (2001) Principles of monochromatic data collection. In: Rossmann MG, Arnold E (eds) International tables for crystallography, vol. F, pp. 177–195
2. Dauter Z (2010) Carrying out an optimal experiment. Acta Crystallogr D66:389–392
3. Popov AN, Bourenkov GP (2003) Choice of data-collection parameters based on statistic modelling. Acta Crystallogr D59:1145–1153
4. Bourenkov GP, Popov AN (2006) A quantitative approach to data-collection strategies. Acta Crystallogr D62:58–64
5. Bourenkov GP, Popov AN (2010) Optimization of data collection taking radiation damage into account. Acta Crystallogr D66:409–419
6. Leal RM, Bourenkov GP, Svensson O, Spruce D, Guijarro M, Popov AN (2011) Experimental procedure for the characterization of radiation damage in macromolecular crystals. J Synchrotron Radiat 18:381–386
7. Arndt UW, Wonacott AJ (1977) The rotation method in crystallography. North Holland, Amsterdam
8. Pflugrath JW (1999) The finer things in X-ray diffraction data collection. Acta Crystallogr D55:1718–1725
9. Yeates TO (1997) Detecting and overcoming crystal twinning. Methods Enzymol 276: 344–358
10. Henderson R (1990) Cryo-protection of protein crystals against radiation damage in electron and X-ray diffraction. Proc Roy Soc London B241:608
11. Owen LO, Rudino-Pinera E, Garman EF (2006) Experimental determination of the radiation dose limit for cryocooled protein crystals. Proc Natl Acad Sci U S A 103:4912–4917
12. Garman EF (2010) Radiation damage in macromolecular crystallography: what is it and why should we care? Acta Crystallogr D66: 339–351
13. Winter G (2010) xia2: an expert system for macromolecular crystallography data reduction. J Appl Cryst 43:186–190
14. Incardona M-F, Bourenkov GP, Levik K, Pieritz RA, Popov AN, Svensson O (2009) EDNA: a framework for plugin-based applications applied to X-ray experiment online data analysis. J Synchrotron Radiat 16: 872–879
15. Mueller Dieckmann C, Panjikar S, Tucker PA, Weiss MS (2005) On the routine use of soft X-rays in macromolecular crystallography. Part III. The optimal data collection wavelength. Acta Crystallogr D61:1263–1272
16. Evans G, Pettifer R (2001) CHOOCH: a program for deriving anomalous-scattering factors from X-ray fluorescence spectra. J Appl Crystallogr 34:82–86
17. Sun PD, Radaev S, Kattah M (2002) Generating isomorphous heavy-atom derivatives by a quick-soak method. Part I: test cases. Acta Crystallogr D58:1092–1098
18. Dauter Z, Dauter M, Rajashankar KR (2000) Novel approach to phasing proteins: derivatization by short cryo-soaking with halides. Acta Crystallogr D56:232–237
19. Dauter Z (2005) Use of polynuclear metal clusters in protein crystallography. Compt Rend Chim 8:1808–1814
20. Diederichs K, Karplus PA (1997) Improved R-factor for diffraction data analysis in macromolecular crystallography. Nat Struct Biol 4:269–275

21. Weiss MS, Hilgenfeld R (1997) On the use of merging R factor as a quality indicator for X-ray data. J Appl Crystallogr 30:203–205
22. Karplus PA, Diederichs K (2012) Linking crystallographic model and data quality. Science 336:1030–1033
23. Schneider TR, Sheldrick GM (2002) Substructure solution with SHELXD. Acta Crystallogr D58:1772–1779
24. Dauter Z (1999) Data collection strategies. Acta Crystallogr D55:1703–1717