International Tables for Crystallography (2006). Vol. F, Section 9.1.13, pp. 192–194.

energy from the absorbed photons may initially result in the disruption of chemical bonds, before being eventually dissipated as thermal energy. For well ordered small-molecule crystals the lattice is close packed and the effects arising from the absorbed photons are restricted to the immediate environment of the absorption event, so-called primary damage. Only when a substantial fraction of the crystal has been affected do cooperative effects set in.

In contrast, roughly 50% of a macromolecular crystal is disordered aqueous solvent (Matthews, 1968). At room temperature this allows a secondary mechanism of radiation damage, resulting from diffusion of radicals and ions produced at the primary absorption site that affects chemical moieties at positions remote from this site. The details of this process remain poorly understood but are related to the extremely damaging effects of X-rays on biological tissue. A consequence of this damage is that degradation of the crystal order continues even after the irradiation is stopped or interrupted. For collection of data at room temperature from protein crystals mounted in capillaries, secondary damage contributes significantly to the rate of deterioration of the diffraction pattern. One of the gains of the early applications of SR was that it allowed recording of data to proceed ahead of the effects of secondary damage, increasing the effective, if not the absolute, lifetime of the crystal in the X-ray beam. An experiment often required several crystals, all of which showed the effects of temporal decay in their recorded intensities, which needed to be merged to provide complete data.

9.1.12.2. *Cryogenic freezing*

In the early 1990s, the introduction of protein-data collection at cryogenic temperatures, using so-called flash freezing, was a major breakthrough (Garman & Schneider, 1997; Rodgers, 1997). Flash-frozen crystals largely prevented the effects of secondary damage. On the X-ray sources then available, it was in most cases possible to record complete data from a single sample without significant degradation of the diffraction, enormously simplifying the strategy of data collection and merging.

The techniques of macromolecular cryocrystallography have advanced so rapidly that almost all data are currently collected from frozen samples. The key aspects of flash freezing are addressed in Part 10. The prolonged life of the sample and modest rates of data acquisition, even at second-generation SR sources with imaging plates, allowed enough time for careful analysis of the initial images and optimization of the strategy.

A second major advantage of cryogenic freezing is that it allows crystals to be reused after initial data have been recorded. Two examples show the usefulness of this approach. Firstly, when screening the binding of heavy atoms for phase determination or ligands for complex formation, data can first be recorded to the minimum resolution needed to determine whether the binding is successful. Secondly, a series of frozen crystals can be screened for their degree of order in the home laboratory, and the best stored and retained for subsequent improved collection either in the home laboratory or at a synchrotron site. The ability to transport frozen crystals has proved invaluable in this respect, and leads to optimal use of synchrotron resources.

9.1.12.3. *Ultra high intensity SR sources*

The advent of third-generation SR sources and insertion devices has led to X-ray beams of unprecedented intensity, for example at the ESRF or APS. At the time of writing, the first of these beamlines have only recently been commissioned and it is hard to give a precise evaluation of their implications for data-collection strategy. Hence the experience to date is somewhat anecdotal and is not based on published reports.

The speed of data collection can be of the order of 1 second per 1° rotation. In association with CCD detectors able to read out images within a few seconds, this means that a complete data set can be obtained in a few minutes. At first sight this would seem to have solved the problem of macromolecular data collection, as such speeds should allow recording of highly redundant accurate data to the highest resolution in a tractable time. However, with these ultra high intensities it appears that a new element of damage can occur. The useful active exposure lifetime of typical crystals seems to be around five minutes, with substantial degradation of the diffraction pattern ensuing even for cryogenically frozen crystals. This may be a limitation of the rate at which heat resulting from the absorption of photons can be dissipated, with local heat gradients perhaps being the factor responsible for the disruption of the crystal order.

This effect suggests that adopting strategies for choosing the optimal starting point of rotation in the minimal total rotation approach for complete data may once more be vital. Using current software this can be achieved in a matter of minutes. It is worth sacrificing this time for the sake of data quality.

9.1.13. Relating data collection to the problem in hand

The data-collection protocol should be matched to the purposes for which the data are to be used. Different applications present a range of different needs, requiring the intensities (structure-factor amplitudes) to be exploited in different ways. In this section a representative set of applications is outlined in terms of how the tactics and strategies of data collection can vary.

9.1.13.1. *Isomorphous-anomalous derivatives*

The phasing of proteins by isomorphous replacement requires the collection of data from crystals of one or more heavy-atom derivatives of the protein that are isomorphous to the parent native crystal. Preparation of derivatives involves either soaking of native crystals in the heavy-atom solution or co-crystallization with the heavy-atom reagent (Part 12). Data collection can be split into two parts. The first step is to establish whether a potential derivative is isomorphous and contains the expected heavy atoms. The second is to collect the data on this derivative to provide the necessary phase information for the native structure factors. The problems of how to utilize the phase information are addressed in Part 12. Here, strategies applicable to the two steps are described.

Screening of derivatives can be carried out by collecting data to the resolution limits of the crystals. This can consume substantial data-collection resources and lead to irrelevant data that are not from isomorphous crystals or do not contain the anticipated heavy-atom signal. It is preferable to record the minimum data sufficient to identify a potential derivative in order to save time and resources, as many samples may need to be screened. A minimal strategy can exploit some or all of the following protocols:

(1) An essentially complete native-data reference set should be available, although not necessarily to the ultimate resolution limit.

(2) Preparation of a set of crystals with a selected set of potential heavy atoms, the number depending on crystal availability.

(3) Collection of a small number of images from each potential derivative crystal, ideally on the home-laboratory rotating-anode source or an SR beamline if necessary. These data can be recorded to a low resolution: in principle 4 Å or less should be enough. The resulting partial derivative data are scaled with the complete native set. The fractional isomorphous difference can be evaluated easily and compared with the expected agreement with the native data. In general, values less than 10% suggest that the heavy atom is not bound. Values higher than about 30% suggest an unacceptable level of non-isomorphism. Intermediate values suggest, but do not

guarantee, that the derivative is worth pursuing. Normal probability plots can be helpful in this respect (Howell & Smith, 1992).

(4) Given a positive result from point (3), complete data may be recorded on the same or an equivalent crystal. Again, it may be useful to record data to low resolution in the first instance. 4 Å resolution is again quite sufficient to solve the structure of a heavy-atom constellation using direct or Patterson methods, allowing the more complete characterization of the potential derivative.

(5) If the compound proves to be a useful derivative, data can then be recorded to higher resolution for the computation of phase information. It may not be appropriate to record data to the highest resolution as for the native protein. In this context, the strength of the data is of primary importance, and relatively weak data at high resolution may be less relevant.

Some practical points are highly relevant here. The ability to store and reuse frozen crystals means that potential derivatives can first be screened at the lowest possible resolution, and the crystal preserved and used later only if the derivative proves to provide useful phase information. The final resolution for data collection will then depend on the degree of isomorphism. The wavelength, if tunable, should be set to a value just below the absorption edge in order to maximize the anomalous signal. The redundancy can also play an important role, as it is useful to have a large number of independent measurements so that outliers in the native or derivative data can be excluded, as these can cause major problems in either the Patterson or direct-methods approaches for locating the heavy atom (Part 12).

9.1.13.2. *Anomalous scattering, MAD and SAD*

The requirements for collecting data with an intrinsically weak anomalous signal are several. As with the isomorphous measurements in the previous section, the highest possible resolution may not be the primary consideration. Here the emphasis lies in data quality, as the measurement of very small differences in macromolecular amplitudes, which are already in themselves relatively weak, is required. Important considerations include the following.

(1) Optimization of the wavelength, particularly for MAD experiments.

(2) Ensuring that the anomalous data are complete in terms of all possible Bijvoet pairs. This is not always addressed by the currently available data-processing software.

(3) High redundancy of measurements significantly enhances the quality of the signal, as this provides effective averaging of errors and allows the rejection of statistical outliers. The latter is especially important for direct-methods solution of the anomalous-scattering constellation.

For MAD experiments (Hendrickson, 1991; Smith, 1991), which can only be carried out at SR sites, the optimum number of wavelengths at which data should be recorded remains unclear. The minimum is one (SAD) and the conventional wisdom is that four are optimal. Given finite beam time, the trade-off is between measuring with limited redundancy at several wavelengths as against higher redundancy at a smaller number of wavelengths. The jury is still out on this one.

Single-wavelength anomalous dispersion (SAD) represents the limiting case. All data are recorded at one wavelength, reducing the requirement for fine monochromatization and for fine tunability and stability. Now quality, especially in the form of redundancy, is the dominating factor since all phasing is based purely on a single anomalous difference for each reflection.

9.1.13.3. *Molecular replacement*

For the initial data required for molecular replacement (MR), high resolution is not essential. Firstly, the method depends on homologous models that are usually only an imperfect representation of the structure under investigation and hence high-resolution data cannot be accurately modelled, and will only introduce noise into the analysis. Secondly, the rotation function, the first step in MR, is based on the representation of the Patterson function in terms of spherical harmonics, which is limited in its accuracy.

In contrast, it is essential for MR applications that the most intense low-resolution terms are measured. The lack of such reflections strongly affects the rotation- and translation-function computations, as the functions are based on Patterson syntheses involving the square of the structure-factor amplitudes, and are dominated by the largest terms. Elimination of the strongest few per cent of the low-resolution data may well prevent a successful solution by MR.

However, for refinement of structures solved by MR, it is essential that data be recorded to a resolution sufficient to allow escape from the phase bias introduced by the model.

9.1.13.4. *Definitive data on relevant biological structures*

Here it is intended to include all structures that benefit from the highest accuracy in their atomic coordinates to shed light on the details of their biological function. These may include substrate or inhibitor complexes and mutants if the analysis requires the full potential of X-ray crystallography. Many of these will not diffract to atomic resolution; nevertheless, all steps in a detailed crystal structure analysis are made simpler as the resolution and quality of the data are increased. This includes the solution of the phase problem, interpretation of the electron-density maps and the refinement of the model.

The most appropriate strategy for data collection involves decisions based on a complex and mutually dependent set of parameters including:

(1) Crystal quality and availability. If only one crystal is available, the choices are limited. If many are available, then some experimentation is recommended to select a high-quality sample.

(2) Cryogenic freezing. This has become *de rigueur* for the modern protein crystallographer. In many cases it allows collection of data from a single crystal. If appropriate cryogenic freezing conditions cannot be established, making it necessary to record room-temperature data, this can affect strategy-making dramatically, in that several crystals might well be required to achieve the target resolution and completeness.

(3) X-ray source and detector. The availability of these again places restrictions on the experiments which are tractable. An SR source will always provide better data, but has logistical problems of availability and access. For some problems, SR becomes *sine qua non* and a rotating anode is just insufficient. These include the use of MAD techniques, very small crystals, large and complex structures with large unit cells such as viruses, and where atomic resolution data are needed.

(4) Overall data-collection time allocated. This has an obvious overlap with point (3). In particular, if SR is to be used later, then the resolution limit on the home source may be modest. If SR is not likely to be employed, then a higher resolution may be aimed for, requiring more time, and again dependent on the pressure on local resources.

Whatever the resource, it is good to define a strategy that will provide high completeness of the unique amplitudes at the highest resolution, with the realization that there is some conflict between these two requirements.

9.1.13.5. *A series of mutant or complex structures*

The detailed geometry of the molecule is already known and the rather general effects of ligand binding or mutation can be initially identified at a relatively modest resolution and completeness. As with heavy-atom screening, it is often advisable to check that the desired complex or structural modification has been achieved by first recording data at low resolution.

However, if the analysis then proves to be of real chemical interest, with a need for accurate definition of structural features, the data should be subsequently extended in resolution and quality. As with the identification of isomorphous derivatives, this approach has benefited greatly from cryogenic freezing, where the sample can be screened at low resolution and then preserved for subsequent use.

9.1.13.6. *Atomic resolution applications*

As for MAD data, the needs for atomic resolution data are extreme, but rather different in nature. Atomic resolution refinement is addressed in Chapter 18.4. Suffice it to say that by atomic resolution it is meant that meaningful experimental data extend close to 1 Å resolution. There are two principal reasons for recording such data. Firstly, they allow the refinement of a full anisotropic atomic model, leading to a more complete description of subtle structural features. Secondly, direct methods of phasing are largely dependent upon the principle of atomicity.

The problems likely to be faced include:

(1) The high contrast in intensities between the low- and high-angle reflections. This may be much larger than the dynamic range of the detector. If exposure times are long enough to give good counting statistics at high resolution, then the low-resolution spots will be saturated. The solution is to use more than one pass with different effective times.

(2) The overall exposure time is often considerable and substantial radiation damage may finally result. The completeness of the low-resolution data is crucial, and it is recommended to collect the low-resolution pass first as the time taken for this is relatively small.

(3) The close spacing between adjacent spots within the lunes on the detector, dependent on the cell dimensions. The only aid is to use fine collimation.

(4) The overlap of adjacent lunes at high diffraction angle, especially if a long cell axis lies along the beam direction. Using an alternative mount of the crystal is the simplest solution. Otherwise the rotation range per image must be reduced, increasing the number of exposures. This is again a problem with slow read-out detectors.

(5) For direct-methods applications, a liberal judgement of resolution limit should be adopted. Even a small percentage of meaningful reflections in the outer shells can assist the phasing. These weak shells can be rejected or given appropriate low weights in the refinement. The strong, low-resolution terms are vital for direct methods.

9.1.14. The importance of low-resolution data

The low-resolution terms define the overall shape of the object irradiated in the diffraction experiment. Omission of the low-resolution reflections, especially those with high amplitude, considerably degrades the contrast between the major features of the object and its surroundings. For a macromolecule, this means that the contrast between it and the envelope of the disordered aqueous solvent is diminished and, furthermore, the continuity of structural features along the polymeric chain may be lost. Refinement and analysis of macromolecules at all resolutions, be it high or low, involves the inspection of electron-density syntheses. These can be interpreted visually, on a graphics station, or interpreted automatically with a variety of software. In all of these, at all resolutions, the importance of the low-resolution terms is crucial. A special problem is in the interpretation of the partially ordered solvent interface. The biological activity of most enzymes and ligand-binding proteins is located precisely at this interface, and for a true structural understanding of how they function this region should be optimally defined. This is seriously impaired by the absence of the strong, low-resolution terms. The problems become more severe as the upper resolution limit of the analysis becomes poorer. Thus at 1 Å resolution, the omission of the 7 Å data shell will have less effect compared with a 3 Å analysis – but remember that ideally, no low-resolution data should be omitted!

In some phasing procedures, the presence of complete, especially high-intensity, low-resolution, data is even more crucial. The big, low-resolution amplitudes dominate the Patterson function and methods based on the Patterson function are therefore especially sensitive. This encompasses one of the major techniques of phase determination for macromolecules: molecular replacement. Direct methods of phase determination utilize normalized structure factors and predominantly exploit those of high amplitude. The relations between the phases of those reflections with high amplitudes, such as the classical triple-product relationship, are strongest and most abundant for reflections with low Miller indices, hence at low resolution.

The importance of the low-resolution reflections in terms of geometric and qualitative context cannot be overemphasized.

9.1.15. Data quality over the whole resolution range

It is not possible to judge data quality from a single global parameter, especially $R_{\rm merge}$, not even from the overall $I/\sigma(I)$ ratio. Such a parameter may totally neglect problems such as the omission of all low-resolution terms due to detector saturation. A set of key parameters including $I/\sigma(I)$, $R_{\rm merge}$, percentage completeness, redundancy of measurements and number of overloaded high-intensity measurements must be tabulated in a series of resolution shells. This information should be assessed during data collection to guide the experimenter in the optimization of such parameters as exposure time, attainable resolution and required redundancy. As stated in Section 9.1.13, the requirements will vary with the application.

The effect of sample decay also requires such tables. The X-ray intensities decay more rapidly at high angle than at low, and consideration of this effect requires knowledge of the relative B values that need to be applied to the individual images during data scaling. An often subjective decision will need to be made regarding at what stage the decay is sufficiently high that further images should be ignored. The effects of damage are likely to be systematic rather than just random, and cannot be totally compensated for by scaling. This remains true even for cryogenically frozen crystals, especially with ultra bright synchrotron sources.

Following an earlier recommendation by the IUCr Commission on Biological Molecules (Baker *et al.*, 1996), this tabulated information, as a function of resolution, should be deposited with the data and the final model coordinates in the Protein Data Bank. Only then is it possible to have a true record of the experiment and for users of the database to judge the correctness and information content of a structural analysis.

9.1.16. Final remarks

Optimal strategies for data collection are dependent on a number of factors. The alternative data-collection facilities to which access is potentially available, how long it takes to gain access and the overall time allocated all place restraints on the planning of the experiment. In view of this, it is not possible to provide absolute rules for optimal strategies.

references