the FEBS Journal

REVIEW ARTICLE

Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination

Alexander Wlodawer[1], Wladek Minor[2,3,4,5], Zbigniew Dauter[6] and Mariusz Jaskolski[7]

1 Protein Structure Section, Macromolecular Crystallography Laboratory, NCI at Frederick, Frederick, MD, USA
2 Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA
3 Midwest Center for Structural Genomics, USA
4 New York Structural Genomics Consortium, USA
5 Center for Structural Genomics of Infectious Diseases, USA
6 Synchrotron Radiation Research Section, Macromolecular Crystallography Laboratory, NCI, Argonne National Laboratory, IL, USA
7 Department of Crystallography, Faculty of Chemistry, A. Mickiewicz University and Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland

This paper is a tribute to Professor Richard Perham in appreciation of his long and dedicated service as Editor-in-Chief of the *FEBS Journal*.

Keywords
data collection and processing; electron density maps; protein crystallography; structure refinement; structure solution; structure quality; structure validation

Correspondence
A. Wlodawer, Protein Structure Section, Macromolecular Crystallography Laboratory, NCI at Frederick, Frederick, MD 21702, USA
Fax: +1 301 846 6322
Tel: +1 301 846 5036
E-mail: wlodawer@nih.gov

(Received 16 July 2013, revised 12 August 2013, accepted 20 August 2013)

doi:10.1111/febs.12495

The number of macromolecular structures deposited in the Protein Data Bank now approaches 100 000, with the vast majority of them determined by crystallographic methods. Thousands of papers describing such structures have been published in the scientific literature, and 20 Nobel Prizes in chemistry or medicine have been awarded for discoveries based on macromolecular crystallography. New hardware and software tools have made crystallography appear to be an almost routine (but still far from being analytical) technique and many structures are now being determined by scientists with very limited experience in the practical aspects of the field. However, this apparent ease is sometimes illusory and proper procedures need to be followed to maintain high standards of structure quality. In addition, many noncrystallographers may have problems with the critical evaluation and interpretation of structural results published in the scientific literature. The present review provides an outline of the technical aspects of crystallography for less experienced practitioners, as well as information that might be useful for users of macromolecular structures, aiming to show them how to interpret (but not overinterpret) the information present in the coordinate files and in their description. A discussion of the extent of information that can be gleaned from the atomic coordinates of structures solved at different resolution is provided, as well as problems and pitfalls encountered in structure determination and interpretation.

Abbreviations
3D, three-dimensional; ADP, atomic displacement parameter; BPTI, bovine pancreatic trypsin inhibitor; CCD, charge-coupled device; DM, direct methods; MIR, multiple isomorphous replacement; MR, molecular replacement; PAD, pixel array detector; PDB, Protein Data Bank; RSCC, real-space correlation coefficient; RSR, real-space residual; SAD/MAD, single/multiple anomalous diffraction; SG, structural genomics; TLS, translational/librational/screw; XFEL, X-ray free electron laser.

Introduction

Protein crystallography is a branch of science that is now considered to be quite mature, with the unintended consequence that fewer and fewer scientists are actually trained in its application. The users of macromolecular structures often know even less about how far to trust the published information. Thus, some time ago, we published a didactic review that aimed to remedy this situation [1]. Although the title was 'Protein crystallography for non-crystallographers ...', the review contained material that could be useful also to less-experienced practitioners of this field. We have learned that that review has been used in many academic and commercial structural biology laboratories as a manual for new members. During the more than 5 years since the original publication, the number of structures deposited in the Protein Data Bank (PDB) [2,3] has approximately doubled, and many of the new structures have been determined by scientists who are principally active in other areas of structural biology, or even of just biology in general. Although the availability of modern instruments and software tools makes structure determination appear almost routine, this is quite often not the case. Hence, the present update of the original review is directed not only to the users of macromolecular structures, but also to some of the younger providers of structural data. In the spirit of full disclosure and to prevent possible flagging of the present review as self-plagiarism, we would like to clearly state that it is based, to a certain extent, on the material previously published in this journal [1], although, here, we address the substantial progress that was achieved in structure determination/analysis during last 5 years. An important addition to the previously published material is a glossary presented at the end of the present review, in which we define the terms that need to be known by all practicing crystallographers, with some of them being also useful for noncrystallographers who utilize structural data. We treat the preparation of the previous and the present publication as fulfillment of our mission as educators, a mission that is particularly pressing in the light of the almost complete disappearance of crystallography courses from the university curricula worldwide. Parenthetically, this ban appears to be totally irrational in view of the past, present and foreseeable success of this discipline, a fact that is even celebrated by the United Nations in their declaration of 2014 as the International Year of Crystallography (IYCr2014).

Macromolecular crystallography has changed in a very major way during the more than half-century since the first protein structure (of myoglobin at 6 Å resolution) [4] was published. Although only seven protein structures were included in the PDB when it was established in 1971, the number now exceeds 93 000, with more than 80 000 of them determined by crystallography. The pace of structure determination has accelerated during the last quarter century owing to automation of protein production and crystallization, the routine availability of very powerful synchrotron X-ray sources, as well as the introduction of sophisticated new algorithms and computer software for diffraction data collection, structure solution, refinement and presentation. Of particular importance are structural genomics (SG) efforts conducted in a number of centers worldwide, which can be credited with more than 12 000 deposited crystal structures as of June 2013 (W. Minor, unpublished data). Although the total number of unique protein folds that can be found in nature is still under debate [5] and the structures of many proteins, especially those integral to cell membranes, are still unknown, the gaps in our knowledge are being filled quite rapidly.

Protein crystallography was once an arcane technique accessible only to well-trained scientists and its initial development was a real tour-de-force. It is not surprising that determination of the structure of haemoglobin took more than 20 years because all of the appropriate methodology had to be invented [6]. It is less obvious that, even later, some structures of comparatively simple proteins took many years to solve. An example may be provided by nerve growth factor, a dimer of two chains, each comprising 118 amino acid residues. Although the first crystallization report was published in 1975 [7] and work was continued uninterrupted in subsequent years, the structure itself was determined only in 1991 [8]. Not surprisingly, determination of the structures of some complex molecular machines was also a long and laborious process. An example of a monumental work that led to a Nobel Prize is the determination of the structure of the ribosome. Although microcrystals of ribosomes were isolated as early as 1970 [9] and single crystals were grown a decade later [10], it took another two decades of extensive efforts in several laboratories to provide the first atomic models [11–13]. More routine structures can now be determined very quickly, even in hours, although sometimes the resulting publication may show the limited experience of the investigators in solving and interpreting the structures of macromolecules. Although several rigorous [14,15] or simpler [16–18] textbooks describing protein crystallography have been published in the years subsequent to the appearance of the iconic book by Blundell and Johnson [19], it appears that the information contained in them is not always fully utilized. Another very important reference for all crystallographers are the monumental multivolume *International Tables for Crystallography* [20]. Volume A, in particular, contains a comprehensive description of crystal symmetry, and volume F is dedicated to protein crystallography. Because the mathematical treatment presented in volume A is rather advanced, it is possible to refer to a simplified description of how to interpret the contents [21]. It may also not be completely clear to noncrystallographers regarding how to interpret some other technical aspects of crystallographic data, although it has always been obvious that 'macromolecular structure matters' [22]. In the present review, we address all of these problems.

When the first protein structures were published, they were rather crude, although even then much could be learned from them about how the molecules look and function. Ever since structure refinement became routine, their quality has vastly improved, although an assessment of the quality of macromolecular structures corrected for technical difficulty, novelty, size and resolution [23] has concluded that, on average, the quality of protein structures has been quite constant over approximately 35 years. However, it is now technically possible to continually re-refine the structures deposited in the PDB using the original experimental data but with the ever improving computational tools, although not much further improvement in structure quality is expected [24]. We will discuss some aspects of such efforts as well.

Preparation and crystallization of macromolecules

Studies of various protein crystals predate the first diffraction experiments; for example, 600 microscopic photographs of the crystals of haemoglobin from the blood of different animals, as well as their detailed description, were published over a century ago [25]. Of course, these were purely phenomenological descriptions that did not provide any information about the structure of the proteins themselves; the first structures of myoglobin and haemoglobin became available only half a century later [4,6]. In those cases, as well as in all early crystallographic studies of proteins and nucleic acids, the macromolecules were isolated from their natural sources by a variety of biochemical procedures. Structural studies of proteins directly isolated from such sources are now an exception rather than a rule because the vast majority of target macromolecules are produced through genetic engineering of bacteria, primarily *Escherichia coli*, or in a variety of eukaryotic cells.

The crystallization of biological macromolecules has been described in great detail in classic textbooks [26–28], as well as in many research papers. Although, early on, each crystallographer approached the problem in an individual way, the procedures have been largely standardized, especially as a result of the availability of crystallization kits, as well as robots for the preparation of solutions, setting up crystallizations and even for seeding, etc. The earliest batch methods that involved simply mixing protein solutions with precipitants are still in use in special cases or in smaller laboratories, although their significance for routine crystallizations is declining as a result of the domination of high-throughput approaches. Equilibration involving diffusion through membranes is also rare because it does not lend itself to automation. The most common modes of equilibrating protein solutions with precipitants involve vapour diffusion, either in a sitting-drop or hanging-drop mode. For practical details of the procedures, we would recommend the textbooks mentioned above, as well as instructions found on the web pages of the suppliers of crystallization kits.

The availability of crystallization robots and the miniaturization of the crystallization apparatus led to a very significant decrease in the amounts of protein required for setting up a broad preliminary screen. Although the rule-of-thumb used to be that approximately 10 mg of pure protein was needed, even as little as 1 mg may now be sufficient for investigating a very wide range of crystallization conditions. However, the crystallization of a particular protein is still sometimes a hit-and-miss affair, and a number of salvage procedures that involve changing the organism from which a protein is derived, trimming parts of the polypeptide chain by controlled proteolysis, directed mutation of surface residues and utilization of crystallization chaperones, etc., are sometimes needed. The currently used procedures have been frequently described in reviews, such as that by Bukowska and Grütter [29].

It should be noted that there is no perfect correlation between the appearance of a crystal under the microscope and its ability to diffract X-rays. For example, apparently perfect crystals of *Medicago truncatula* serine/threonine protein kinase (Fig. 1A) did not diffract at all and decent looking crystals of survivin B from *Xenopus laevis* (Fig. 1B) did not diffract beyond 10 Å (Fig. 1E), whereas a very odd-looking crystal of a Z-DNA dodecamer (Fig. 1C) provided high-quality data at ultrahigh (0.75 Å) resolution (Fig. 1D).

Particularly difficult crystallization-related problems are presented by integral membrane proteins. Because of difficulties in crystallizing such proteins, their structures represent only a small fraction of the coordinates in the PDB. However, the importance of structural knowledge of integral membrane proteins for the understanding of biologically important processes cannot be

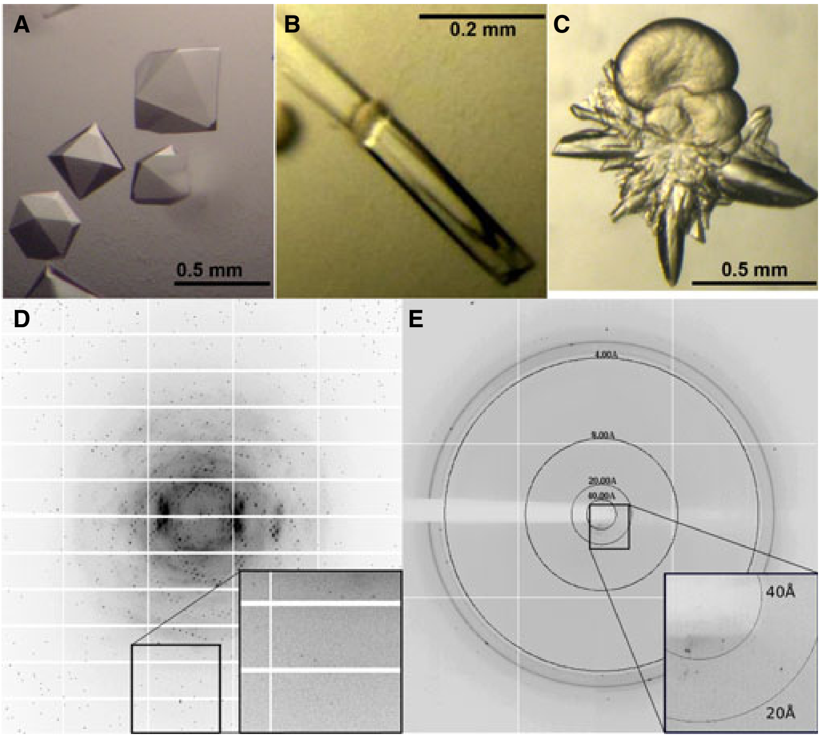

Fig. 1. A challenge: try to match the crystals with their diffraction patterns. Would you be able to match two out of the three crystals shown in (A), (B) and (C) with the X-ray diffraction patterns in (D) and (E)? The answer: the best diffraction pattern (D) was recorded for the ugliest specimen (C). The good looking crystal shown in (B) gave very poor diffraction (E), and the perfect looking crystals in (A) gave no diffraction at all (not shown). (A) Crystals of *M. truncatula* serine/threonine protein kinase. (B) Crystal of survivin B from *X. laevis*. (C) Crystal of a synthetic Z-DNA dodecamer. (D) Diffraction image taken from the top part of the crystal of Z-DNA dodecamer shown in (D). The data (to 0.75 Å resolution) were obtained with a PILATUS detector at the NE-CAT beamline of the Advanced Photon Source (Argonne National Laboratory, IL, USA). (E) Diffraction image of survivin B taken for the crystal shown in (B) with an ADSC Quantum315 detector at the SBC-CAT beamline of the Advanced Photon Source. Only a few weak low-resolution reflections can be seen in the inset. The ring beyond the 4 Å mark is a result of ice and indicates problems with cryo-cooling of this crystal.

overemphasized, as shown by the award of several Nobel Prizes for work that involved crystallographic studies by at least one of the recipients. Deisenhofer *et al.* [30] determined the structure of the photosynthetic reaction center, Jiang *et al.* [31,32] studied water and ion channels and Rasmussen *et al.* [33] investigated structural and biological properties of G-protein-coupled receptors. The latter studies, in particular, required very extensive modification of the receptors utilizing monoclonal antibodies, nanobodies and fusion with T4 lysozyme [34–36]. Useful crystals could not be obtained without such modifications. The question always remains regarding how these modifications affect our biological analysis of macromolecular structures.

X-ray sources used in crystallographic experiments

Macromolecular crystallography relies almost exclusively on the scattering of X-rays by the electrons in the molecules constituting the investigated sample. Scattering of particles, such as neutrons or electrons, is also used to investigate macromolecular crystals, although only a small fraction (< 0.1%) of the published macromolecular structures have been determined this way and they are not discussed here further. It should be emphasized, however, that those methods, although experimentally much more demanding, provide extremely valuable information. Neutron scattering, for example, informs about the coordinates of hydrogen atoms (often crucial to the understanding of the functioning of macromolecules), which are very rarely directly located by the X-ray experiments (owing to their minute contribution to the electron cloud) [37,38].

The sources of X-rays used for crystallographic experiments and the methods of their detection have undergone dramatic changes and improvements during the more than a century since Röntgen's discovery. For most of that period, X-rays were generated in various types of vacuum tubes, in which highly accelerated electrons were bombarding anode targets made of metals such as copper (Cu) or molybdenum (Mo), leading to the emission of characteristic X-rays with wavelengths dependent on the anode material, superimposed on a background of white radiation. Although Mo anodes, generating X-rays with a wavelength of $\lambda = 0.7107$ Å, were traditionally used for data collection from crystals of small molecules, macromolecular crystallographers have usually utilized X-rays generated with Cu anodes (wavelength of $\lambda = 1.5418$ Å). Because the X-ray beam generated with this technique is not monochromatic, it has to be either filtered through material that absorbs the unwanted photons (such as zirconium for Mo radiation or nickel for Cu radiation) or monochromatized using crystals or mirrors. X-ray tubes have undergone major improvements in their design, which replaced sealed tubes with rotating-anode generators, in which the metal target rotation prevents local heat build-up and allows generation of beams with much higher flux.

A major change in the method of X-ray generation took place in the mid-1970s, when the application of synchrotron radiation became practical [39,40]. Although the early use of synchrotron sources (or, more precisely, electron or positron storage rings) by crystallographers was parasitic in nature (those instruments were principally designed for particle physics), purpose-built second- and third-generation storage rings became available in the 1980s and 1990s, and they were vastly improved since then. These instrumental advances increased the available fluxes of X-rays by many orders of magnitude, and allowed easy selection of any wavelength in the approximate range 0.5–3.0 Å, rather than relying on fixed wavelengths of the conventional sources. This added flexibility has led to the universal utilization of anomalous scattering for phase determination (see below).

A new development in the way X-rays are generated is the introduction of X-ray free electron lasers (XFEL) as their source. An XFEL can be considered as a 'cross' between optical lasers (light amplification by stimulated emission of radiation) and a synchrotron. In variance with a classical synchrotron storage ring, in XFEL, the electrons are accelerated in a linear device but, in variance with an optical laser, there is no energy pumping in a closed system; rather, the electrons are accelerated to relativistic velocity by an oscillating microwave field in a long (several kilometres) system of superconducting cavities. The accelerated electrons are then passed on a slalom through a very long (several hundred metres) undulator where they emit electromagnetic radiation. Because the electrons are only somewhat slower than the emitted light, the two waves interact, leading to gradual organization of the electrons into a series of very thin discs in which they behave coherently. In particular, they emit synchronously extremely short and intense flashes of coherent electromagnetic radiation in a process called SASE (self-amplified spontaneous emission). A single accelerator can drive several undulators at the same time. The first XFEL devices [LCLS (Linac Coherent Light Source) at Stanford and the RIKEN XFEL in Japan] are already operational. The European XFEL under construction in Hamburg will have five undulators with ten experimental stations. The flux available from these new sources is again many orders of magnitude higher than what can be obtained even at the best third-generation synchrotron sources. Although practical use of these most modern instruments is still at early testing stages, the first novel protein structure obtained with the use of XFEL at LCLS was recently published [41], and the scientists at RIKEN have shown that the quality of XFEL data collected with their new instrument may be sufficient to detect the anomalous signal of sulfur atoms although, so far, it is not quite adequate for structure determination from such a signal alone [42].

This change in the way X-rays are generated has had a major practical consequence for the operation of crystallographic laboratories. In the past, X-ray generators were found in each laboratory and all data were collected locally. Synchrotron sources, on the other hand, are central facilities, with still only a relatively small number (38 with diffraction capabilities) of such installations operating globally. The originally adopted mode of operation was to preliminarily characterize crystals at a home source, and to travel to a synchrotron to collect the ultimate data used to determine and refine the structures. However, improvements in automation and data transmission technology have led to the current practice, whereby a large number of cryo-cooled crystals are mounted in holders called 'pucks' and sent to a synchrotron facility in dewar shippers. At the beamline, the pucks are placed in robotic devices that allow rapid mounting of the crystals in the X-ray beam. The whole process of data collection can be accomplished by remote computer access or by mail-in crystallography; the latter when data are collected by synchrotron-based staff. Thus, synchrotron-based data collection can be, in principle, a travel-free procedure unless physical presence of the experimenter increases the efficiency of the whole structure determination process or enables, for example, immediate structure–function studies in a feedback loop, whereby the experimenter modifies the research protocol according to structural results generated on-the-fly.

Detectors

Development of better detectors of X-rays also contributed significantly to the improvements of data collection methodology. At first, for more than half a century, the detector of choice was a photographic film (used even for the iconic image of a hand recorded by Röntgen himself). Scintillation counters were later used with X-ray diffractometers, which were superseded by two-dimensional multiwire detectors introduced in the 1980s (their development led to the award of the Nobel Prize in Physics in 1992 to Georges Charpak) and, later, by image plates, charge-coupled devices (CCDs) and pixel array detectors (PADs). Although read-out of a diffraction pattern from a film (including chemical processing, drying and scanning, initially by eye, and later by electronic scanners) would take hours, new detectors can be read-out in a small fraction of a second. With the availability of short-duration pulsed X-ray sources, it is now possible to record X-ray diffraction images with femtosecond exposure [41].

The nature of X-ray diffraction data

Because the highly similar structural motifs that form the individual unit cells are repeated throughout its entire volume in a periodic fashion, a crystal can be treated as a three-dimensional (3D) diffraction grating. As a result, the scattering of X-rays is enormously enhanced in selected directions, and completely extinguished in other directions. This is governed only by the geometry (size and shape) of the crystal unit cell and of the wavelength of the X-rays, which should be of the same range as the interatomic distances (chemical bonds) in molecules. However, the effectiveness of interference of the diffracted rays in each direction, and therefore the intensity of each diffracted ray, depends on the constellation of all atoms within the unit cell. In other words, the crystal structure is encoded in the diffracted X-rays (i.e. the shape and symmetry of the cell define the directions of the diffracted beams, and the locations of all atoms in the cell define their intensities). The larger the unit cell, the more diffracted beams (called 'reflections') can be observed. Moreover, the position of each atom in the crystal structure influences the intensities of all the reflections and, conversely, the intensity of each individual reflection depends on the positions of all atoms in the unit cell. It is therefore not possible to solve only a selected, small part of the crystal structure without modelling the rest of it. This is in contrast to other structural techniques such as NMR or extended X-ray absorption fine structure studies, in which data derived from different parts of a molecule can be separated.

Practical aspects of collecting diffraction data

A diffraction experiment involves measurements of a large number of reflection intensities (an example of a diffraction pattern is provided in Fig. 1D). Because crystals have certain symmetry, some reflections are expected to be equivalent and thus to have identical intensity. The average number of measurements per one individual, symmetrically unique reflection is called redundancy or multiplicity. Because every reflection is measured with a certain degree of random error, the higher the redundancy, the more accurate the final estimation of the averaged reflection intensity. However, the desire for high redundancy, and thus longer and/or more intense exposure of crystals to X-rays, also has a price. The diffraction quality of crystals degrades with cumulative absorbed dose of energy from exposure to high-intensity X-rays, eventually resulting in their severe damage or destruction, as manifested most dramatically by the deterioration of the high resolution limit of the diffraction pattern. Proper application of corrections for sample decay can improve the accuracy of the processed data. Unfortunately, decay as a result of radiation damage is typically associated with specific structural changes, such as breakage of disulfide bonds and decarboxylation of acidic groups, etc., and these effects cannot be easily 'repaired'. Nevertheless, when data redundancy is high, extrapolation to zero absorbed dose may provide some improvement of data quality [43,44]. However, it should always be kept in mind that even the best correction algorithms cannot restore non-existent data. Various types of data collection strategy software exist that can estimate the radiation dose a crystal can withstand and still produce useful data, as well as assess the values of data collection parameters (exposure time, beam attenuation, crystal rotation angles, etc.) that would maximize the amount of information collected from a crystal (or crystals) during accumulation of that dose. Probably the most popular software is BEST [45,46]. However it is surprising that, despite the availability of wonderful strategy software, experimenters still collect many sub-optimal data sets (poorly chosen oscillation range and diffraction covering only a small part of a large detector).

There are other factors that affect the optimal design of a diffraction data collection strategy. Some depend upon the equipment used to perform the experiments: the size and dynamic range of the X-ray detector, the noise associated with the read-out process and the accuracy of spindle axis movement, etc. Other factors are associated with the internal order of the crystal itself and thus are beyond the control of the experimenter (unless, of course, better crystals can be grown): the crystal unit cell parameters, symmetry (crystallographic space group), and mosaicity. Taking all these factors into account requires not only good strategy software, but, most of all, sophisticated experimental protocols. Such protocols need to be optimized for the four types of the experimental data that are routinely collected, for use with: single/multiple anomalous diffraction (SAD/MAD), molecular replacement (MR), ligand searches or structure refinement. An examination of the productivity of synchrotron stations shows that similarly equipped synchrotron beamlines can differ in productivity by an order of magnitude. The difference is even more striking when one examines only SAD/MAD structures, which, by their nature, require higher data quality than MR.

Experimenters do not always realize that what is possible is not necessarily equivalent to what is best for producing the highest quality diffraction data. For example, the new PADs have three important features that are advantageous over the previous generation of CCD detectors: very short time and noise-free read-out, as well as very high dynamic range. The very short read-out time can be used for 'shutterless' data collection and elimination of errors from imperfect synchronization between shutter and spindle axis movement. Unfortunately, the short read-out time encourages collection of images with very small oscillation angles, sometimes as small as 3% of the mosaic spread of a typical crystal. Such a data collection protocol can result in thousands of diffraction images just wasting time and disk space. With only a few X-ray counts even at relatively strong reflection spots, the diffraction pattern may be dominated by noise (elimination of read-out noise does not eliminate other sources of noise, such as incoherent scattering). Thus, in this example, data reduction software must deal with poor data, not as a result of poor crystals but rather a poor data collection strategy. There is no gain in dividing each reflection into more than five images [47] and $\Delta\phi$ values in the range 0.1–0.45° are adequate in most cases. Using the tradition-sanctioned 'standard' value of 1° is almost always wrong for relatively low mosaicity crystals, although it is most often used, evidently for psychological reasons, or maybe because, in the past, scaling software needed the presence of fully recorded reflections. Rotation of 1° may be disastrous for longer unit cells as a result of a large number of overlaps in the medium to high resolution area of the detector.

CCD and PAD detectors may have some inactive or damaged areas or pixels. Figure 1D shows a diffraction image where the inactive regions between the mosaic detector tiles are apparent. Some spurious features (sometimes called 'zingers') or bad individual detector pixels are also clearly visible close to the detector corners, although they do not look so prominent on the zoomed fragment (Fig. 1D, inset). Such bad pixels are ignored by image integration software and do not harm data quality, unless the integration software is unable to correctly recognize and exclude bad pixels during profile fitting or integration of the diffraction spot. In that case, however, the offending reflection may be rejected as an outlier during the data scaling/merging process. Dealing with detector artefacts, crystals with high and/or anisotropic mosaicity, and the pathologies of the experimental system (such as uneven motion of the spindle axis) still presents a challenge for much of the available data reduction software. All of the popular data reduction software [HKL-2000 (Denzo/Scalepack), MOSFLM/AIMLESS and XDS, listed here in cumulative order of their utilization according to the annotation of PDB entries] are able to produce very high quality data given well-diffracting single crystals and correctly performed diffraction experiments. We are not certain how to evaluate the performance of the 'software' NULL, which is consistently acknowledged every year in approximately 400 new PDB entries.

In principle, a macromolecular diffraction experiment is an easy one because only four parameters are controlled by the experimenter (detector–sample distance, wavelength, exposure time and oscillation angle). The difficulty, however, arises from the fact that macromolecular crystals are far from ideal and the experimenter has to decide what trade-offs are best for a particular crystal, type of experiment (MR, SAD/MAD, ligand screening, etc.) and experimental set-up. Moreover, the sample is changing during the experiment as a result of radiation damage. In practice, it is very difficult for a casual user to identify the correlation between crystal and experimental set-up limitations. The best data reduction software allows for simultaneous integration and scaling (and sometimes structure determination) as the diffraction images are collected, which provides almost instantaneous feedback, including information about radiation damage. There is one more, sometimes neglected, role of data reduction software: it allows the beamline personnel to establish the best experimental protocols and identify and correct many equipment malfunctions.

Assessment of the quality of diffraction data

The spread of individual intensities of all symmetry-equivalent reflections, contributing to the same unique reflection, is usually gauged by the residual R_{merge} (sometimes called R_{sym} or R_{int}). The averaging process is known as 'scaling and merging', and its result is a set of unique reflection intensities, each accompanied by standard uncertainty, or estimate of error. Multiple observations of the same reflection provide a means to identify and reject outliers, which may have resulted, for example, from instrumental glitches. However, the number of such rejections should be minimal, a fraction of a percent at most. R_{merge} is defined as $R_{\mathrm{merge}} = \Sigma_{\mathrm{h}}\, \Sigma_{\mathrm{i}}|<I_{\mathrm{h}}> - I_{\mathrm{h,i}}|/\Sigma_{\mathrm{h}}\, \Sigma_{\mathrm{i}}\, I_{\mathrm{h,i}}$ where h enumerates the unique reflections and *i* represents their symmetry-equivalent contributors. Although this indicator has been traditionally reported in most crystallographic papers as an indicator of data quality, it is not perfect [48] because it does not take into account measurement multiplicity. In general, higher multiplicity leads to increased R_{merge} [49], although the availability of more measurements should lead to improved accuracy of the final data set. Thus, more elaborate versions of R_{merge} have been proposed. One of them is R_{meas}, defined as $R_{\mathrm{meas}} = \Sigma_{\mathrm{h}}\, (n_{\mathrm{h}}/n_{\mathrm{h}} - 1)^{1/2}\, \Sigma_{\mathrm{i}}|<I_{\mathrm{h}}> - I_{\mathrm{h,i}}|\, /\Sigma_{\mathrm{h}}\, \Sigma_{\mathrm{i}}\, \mathrm{I}_{\mathrm{h,I}}$, where n_{h} denotes multiplicity [49]. It was shown that the values of R_{meas} for low-redundancy data sets are as high as those for high-redundancy data sets, and that R_{meas} for combining two data sets is close to their individual R_{meas} values in the absence of systematic differences. Thus, the modified indicator of data quality is preferable to the original one but, unfortunately, it is not required by the journals and the PDB. Another index is the precision-indicating merging *R* factor R_{pim}, defined as $R_{\mathrm{pim}} = \Sigma_{\mathrm{h}}\, [1/(/n_{\mathrm{h}} - 1)]^{1/2}\, \Sigma_{\mathrm{i}}|<I_{\mathrm{h}}> - \mathrm{I}_{\mathrm{h,i}}|/\Sigma_{\mathrm{h}}\, \Sigma_{\mathrm{i}}\, I_{\mathrm{h,i}}$ [48,50].

A high-quality set of diffraction data should be characterized by an overall value of R_{merge} (or R_{meas}) of less than 4–5%, although, with well-optimized experimental systems and good-quality crystals, it can be even lower. In our opinion, a value higher than 10% suggests sub-optimal data quality, although data with R_{merge} as high as 20% may still lead to a plausible solution of important/difficult structures. A traditionally accepted indication that the resolution limit has been reached was when the R_{merge} reaches approximately 60% for low-symmetry crystals and even more for high-symmetry crystals, or when the average ratio of reflection intensity to its estimated error, $I/\sigma(I)$ (in some software, this is defined as $<\mathrm{I}/\sigma(\mathrm{I})>$, in other software, it is defined as $<I>/<\sigma(I)>$), drops below approximately 2.0 in the highest resolution shell. However, this cut-off is somewhat arbitrary and, as recently noted, 'an appropriate choice of resolution cutoff is difficult and sometimes seems to be performed mainly to satisfy referees' [51]. It has been recently postulated that these criteria may be too conservative and that even weaker high-resolution data could improve the refined model [52,53]. The same investigators introduced another indicator of data accuracy in the form of $\mathrm{CC}_{1/2}$, which reports the correlation coefficient between reflection intensities in the two halves of a randomly split data set. However, it appears that there is no ultimate and reliable criterion for judging the data resolution limit [53] although it is 'safe' to measure data to higher resolution than suggested by the $I/\sigma(I) = 2.0$ criterion. Inclusion of weak data does not harm the process of structure refinement by contemporary software based on the maximum likelihood principle and, in some cases (e.g. highly anisotropic diffraction), this may be beneficial.

On the other hand, a value of $I/\sigma(I)$ much higher than 2.0 in the outermost shell indicates that the crystal could diffract farther but the resolution of the data was limited by the experimenter or the experimental set-up. This should be avoided because the use of the maximum achievable resolution for refinement not only permits finer structure details to be observed, but also removes possible model bias because a higher resolution improves the data-to-parameter ratio. Unfortunately, in approximately 48% of all PDB deposits that report this value, the $I/\sigma(I)$ in the high resolution bin is higher than 3.0, demonstrating that the full diffraction potential of crystals is often not utilized.

A data set should be as complete as possible (preferably 100%) in all resolution shells. It does not mean, however, that one should artificially truncate high-resolution data only to see high completeness in the last resolution shell. By contrast, all reflections are very precious and should always be included, particularly at high resolution. If completeness in the last resolution shell is really poor, it only means that the realistic resolution limit (to be given, for example, in the publication) will have to be appropriately adjusted. Incomplete low-resolution shells (a serious impediment in case of Patterson-based structure solution, such as MR, or of direct methods) usually result from overloaded strong reflections that cannot be measured because of detector pixels oversaturation. If this is expected to be the case, a quick low-resolution pass of data collection should be collected prior to the long-exposure final pass aimed at accurate measurement of the weak high-resolution data.

The anomalous signal can be judged by criteria analogous to those mentioned above applied to anomalous (Bijvoet or Friedel) differences, i.e. $R_{\mathrm{anom}} = \Sigma_{\mathrm{h}}|\Delta I^{\pm}|/\Sigma_{\mathrm{h}} \langle I \rangle$, Bijvoet ratio $\langle|\Delta F^{\pm}|\rangle/\langle F \rangle$ or ratio of anomalous differences to their uncertainties $\langle|\Delta F^{\pm}|\rangle/\langle\sigma\Delta F\rangle$ (the ± sign indicates a difference between reflections with inversion-center related indices; i.e. *hkl* and $\overline{hkl}$). An indication that the anomalous signal might be useful for phasing is when the correlation between anomalous differences of two randomly split half-data sets is higher than 30% [54].

Common problems with diffraction data

A majority of macromolecular crystal structures are now determined by well-established methods using powerful and user-friendly computer software. However, some cases present problems that require approaches extending beyond the beaten track. Such ‘pathological’ [55] or rather atypical cases involve, for example, crystals characterized by various kinds of pseudosymmetry and/or twinning.

If the molecules in a crystal are arranged in a way that resembles certain symmetry but, in reality, only a subset of symmetry operations is preserved strictly, the determination of the true space group may be difficult. This situation may, for example, occur in cases of translational noncrystallograpic symmetry if multiple copies of the same structural motif present in the asymmetric unit of the crystal have a similar orientation and are related by a parallel shift that is a fraction of the unit cell repeat. This results in abnormalities in the distribution of reflection intensities in the reciprocal space and in difficulties in the assignment of proper symmetry at the data processing stage. Solution of such structures is often possible, although it usually requires a high degree of competence and perseverance in testing various, potentially plausible symmetry interpretations. The unit cell of the crystal of a PR-10/ANS complex (M. Jaskolski and Z. Dauter, unpublished data) illustrates such a case. The diffraction data merged well in tetragonal symmetry, although the structure was successfully solved by MR only after expansion of the data to *P*1 symmetry, which corresponded to as many as 56 individual protein molecules that had to be located by MR. The true symmetry turned out to be monoclinic *C*2 with 28 independent molecules arranged in four parallel columns, each comprising seven molecules with a translation of approximately 1/7 of the cell c parameter.

Frequent pathologies of macromolecular (as well as small-molecule) crystals include various types of twinning [56]. If separate crystalline domains (crystal lattices) within one crystal specimen are in different (but related) orientations such that the unit cells of their lattices overlap, the reflections diffracted by these domains will also overlap. Merohedral twinning can occur when the symmetry of the crystal structure is lower than the full symmetry of the geometrical lattice, as is possible in high-symmetry crystal systems. Pseudomerohedral twinning is possible if the metric of the unit cell is close to that of a higher symmetry system; for example, when a monoclinic unit cell has the β angle very close to 90°. If the twinning is perfect, with the twin fractions (relative volumes of the individual twin domains) close to 0.5, the diffraction data exhibit symmetry that is higher than that of the real structure. Merohedrally-twinned crystals produce data sets characterized by abnormal intensity distribution, which can be used to identify this pathology. Although the initial solution of a crystal structure from twinned data may be relatively more difficult, its further refinement with contemporary refinement software should be straightforward. In essence, the data contain phasing signal mixed from multiple domains and a solution of the individual, unique structure is hampered. This depends on the twinning fraction; if it is small (below 0.1), this effect may not even be noticed.

Several more complicated variants of twinning are possible; for example, when only a subset of reflections overlap in the reciprocal space or when several domains of different but related unit cells occur in one crystalline specimen [55]. It should be noted, however, that it is not correct to use the term ‘twinning’ with reference to diffraction from cracked or disfigured crystals because twinning involves only specific, geometrically-defined relations between crystalline domains.

Some crystals display marked anisotropy of diffraction, where reflection intensities extend in one direction to higher resolution than in other directions. In such cases, the data should be measured up to the resolution limit of the ‘best’ direction, despite the fact that, in other directions, no meaningful intensities will be present and the overall data statistics will be poorer than expected. As noted above, the inclusion of very weak reflections will not harm in a significant way the contemporary phasing and refinement algorithms, which are based on statistically valid principles of e.g. maximum likelihood.

Solution of the phase problem

Each reflection (structure factor) is characterized by its amplitude (obtained as square root of the measured intensity) and phase, although only reflection amplitudes can be obtained from the measured intensities and a diffraction experiment does not provide direct information about reflection phases. However, for the calculation of electron density maps by Fourier synthesis, both the amplitudes and phases of all structure factors are necessary. This constitutes the famous crystallographic ‘phase problem’, the main hurdle in structural crystallography. The phases therefore have to be estimated indirectly. There are three basic types of methods in crystallography to achieve this goal: direct methods, MR and variations of special-atom methods.

Direct methods utilize probabilistic relations between structure factors of certain groups of reflections to estimate their phases, usually by expanding a small set of starting phases, and require that the diffraction data extend at least to atomic resolution, 1.2 Å [57–59]. They are the methods of choice in small-molecule crystallography but are not used to solve large macromolecular structures from the native data alone because the probabilities of phase estimates are inversely proportional to the square-root of the number of atoms (i.e. these are prohibitively small if the number of nonhydrogen atoms in the asymmetric unit exceeds 1000). However, direct methods have been used successfully to solve high-resolution structures of smaller protein molecules; for example, a trypsin inhibitor [60]. More typically, direct methods can be and are routinely used to locate certain ‘special’ atoms, such as heavy or anomalous scatterers, in macromolecular crystals. Because the distances between such atoms are large, even relatively low-resolution data are ‘atomic’ in such cases and direct methods are successfully used for solving heavy-atom or anomalous substructures.

The MR method exploits the easily calculated Fourier transform of reflection intensities (as opposed to structure factors), known as the Patterson function, after Arthur Lindo Patterson who showed that it represents a bunch of all the interatomic vectors (as opposed to all atomic positions). If a suitable atomic model of the unknown crystal structure is available, a model-derived ‘bunch of interatomic vectors’ can be matched with the peaks of the Patterson function, revealing the orientation and location of the model molecule in the unit cell.

MR is currently the most common method for solving protein structures and this success is the consequence of the enormous growth of the contents of the PDB and availability of a large number of various structures suitable as search models. Certainly, not all theoretically possible protein folds are already represented in the PDB but, with the increasing number of deposited structures, the probability of finding a model sufficiently similar for a successful MR search increases as well.

In recent years, several improvements in the theory and practice of MR have been achieved. New algorithms based on proper statistical grounds that exploit Bayesian probability (i.e. maximum likelihood) are implemented in powerful software such as PHASER [61] and MOLREP [62]. Significant progress was made in the process of preparation of improved search models, appropriately modified for a particular sequence. For example, the ROSETTA algorithm [63], which combines MR with structure prediction by modelling, solved several otherwise intractable structures [64]. Moreover, a clever engagement (crowdsourcing) of online players of the protein-folding game FOLDIT (which employs ROSETTA as its workhorse) was used to design a model adequate for successful MR [65,66]. Software pipelines such as BALBES [67] and MRBUMP [68] automatically select the most useful structures from the PDB and sequentially run MR searches with a large number of search models.

The special-atom approach has several variants. It can be based on phasing signal from isomorphous differences between structure amplitudes measured for the native and heavy-atom derivatized crystals or from anomalous differences between centrosymmetrically-related reflections (within Friedel or Bijvoet pairs), resulting from the presence in the sample of anomalously scattering atoms. The classic multiple isomorphous replacement (MIR) method largely lost its importance and popularity in the last decades, even in its version combined with anomalous scattering. This method requires data collected from multiple crystals, derivatized by soaking or co-crystallization with various heavy metal-containing reagents. Such treatment often leads to degradation of the diffraction properties of the crystals or to non-isomorphism, which precludes a successful estimation of the phases. However, soaking the crystal in a solution containing heavy and anomalously scattering elements (such as Pt, Au, Hg, I, Br, etc.) can often lead to an easy and rapid solution of the phase problem by MIR with anomalous scattering, even if the anomalous signal is weak. For large structures, such as viruses or ribosomes, heavy-atom derivatization (often in combination with anomalous signal) by large multi-atom metal clusters [$(Ta_6Br_{12})^{2+}$, $(PW_{12}O_{40})^{3-}$, etc.] provides very powerful phasing at low resolution.

Currently, the method of choice for solving novel macromolecular crystal structures uses phasing based exclusively on the anomalous scattering effect. It requires rather accurately measured diffraction data because the anomalous differences are small (approximately 1–5%) compared to isomorphous differences (approximately 15–25%). However, the availability of stable and tunable synchrotron X-ray sources, sensitive and accurate detectors, and powerful phasing software makes the MAD and the SAD approaches, for which only one crystal is sufficient, the most popular experimental phasing methods. The popularity is also a result of the widespread use of genetic engineering for protein production, which allows easy preparation of recombinant proteins with the anomalously scattering selenium (with an absorption edge at the wavelength of approximately 0.98 Å), introduced into the protein sequence in the form of selenomethionine instead of the native sulfur-containing methionine [69]. Several other anomalous scatterers are also utilized, such as the traditional heavy metals, halides [70], or even sulfur in native proteins [71,72] or phosphorus in nucleic acids [73]. The SAD approach is technically simple, can be executed rapidly and, in principle, requires only one data set collected from one crystal, although data can also be merged from several crystals if radiation damage is severe [74,75]. SAD [76] is responsible for the majority of novel crystals structures deposited in the PDB, including those from high-throughput studies carried out by various SG centers.

In variance with SAD, which relies on a single wavelength (where the anomalous signal is high), MAD [77,78] uses several data sets collected at several wavelengths near the absorption edge, usually at the inflection point of the edge (to maximize the real component of the anomalous correction), at the absorption peak (to maximize the imaginary component) and at a high-energy wavelength remote from the absorption peak. Although MAD requires recording of the X-ray fluorescence spectrum for the selection of the appropriate wavelengths, this requirement is not rigorous for SAD.

Electron density maps and creation of initial models

The primary result of an X-ray diffraction experiment is a map of electron density within the crystal. This electron distribution is usually interpreted in (chemical) terms of individual atoms and molecules, although it is important to realize that the molecular model consisting of individual atoms is already an interpretation of the primary result of the diffraction experiment. Of course, the interpretability of the electron density maps and thus also the accuracy of the atomic models depend on data resolution (i.e. on the number of reflections included in the map computation). Figure 2 shows that a resolution higher than 1 Å permits the confident location of individual atoms, whereas, at a low resolution of approximately 3 Å, it is only possible to locate known fragments, such as protein or nucleic acid residues.

The true representation of electron density is obtained by Fourier summation of $|F_{obs}|$ in combination with the final (i.e. as 'true' as possible) phases. In practice, a model is inspected against a $2F_{obs} - F_{calc}$ map, which (at least in principle) should also reveal errors in the current model because it is a sum of the F_{obs} map and the $F_{obs} - F_{calc}$ difference map. The difference map is also used in its own right and is usually

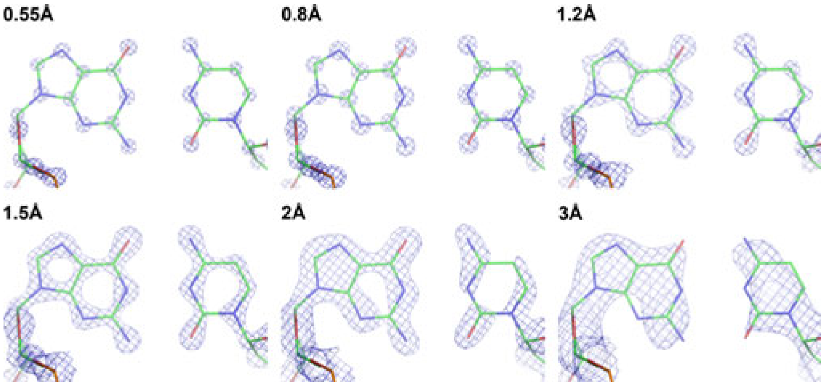

Fig. 2. The appearance of electron density as a function of the resolution of the experimental data. A cytosine–guanine pair from the structure of a Z-DNA hexamer duplex (PDB code: 3P4J) [135] with the (F_{obs}, α_{calc}) maps calculated with different resolution cut-offs at 0.55, 0.8, 1.2, 1,5, 2.0 and 3.0 Å. Although, at the highest resolution of 0.55 Å, there were 75 122 reflections included in map calculation, at 3 Å resolution, only 573 reflections were used.

displayed at two (positive and negative) contour levels. In practice, these maps are computed with statistically weighted coefficients (e.g. $2mF_{obs} - DF_{calc}$ with maximum likelihood-based coefficients) and with phases corresponding to the current model.

Electron density maps should be most properly contoured in e/Å^3 units. However, absolute scaling of electron density is problematic because many of the largest, very low resolution structure factors are missing (i.e. too strong to measure or too close to the primary beam) and, in addition, the map calculation often includes various mathematical tricks (e.g. sharpening). For these reasons, in practice, electron density is usually contoured in the units (known as σ) of the rmsd from the mean electron density level. It should be noted that the σ unit reflects the map noise and will therefore depend on the solvent content (for $2F_{obs} - F_{calc}$ maps) or on the quality of the model itself (for $F_{obs} - F_{calc}$ maps), or even on the region over which the map statistics are calculated. $2F_{obs} - F_{calc}$ maps should be contoured at 1σ or higher and $F_{obs} - F_{calc}$ maps at ± 2.5–3.0σ or higher. Lowering the contour level to visualize at all costs a strongly desired (phantom) feature is deceiving and may have lamentable consequences. Nevertheless, methods to interpret low electron density regions, mostly in terms of introducing multiple models (not unlike the ensembles typical for NMR structures) have been introduced with some success [79,80]. However, the implementation of such composite models, especially at the refinement stage, should be carried out with extreme caution because the incorrect use of such procedures [81] may have unintended and unpleasant consequences (see below).

In small molecule crystallography, it is possible to obtain an accurate atomic structure automatically, without visual comparison of the model with the electron density map. Automatic building of macromolecular models into the initial electron density map can also be accomplished with the help of software such as ARP/WARP [82], RESOLVE [83], BUCCANEER [84] or routines available in COOT [85]. These software packages are often able to automatically produce models of almost complete protein chains at resolutions as low as 3 Å and are now routinely used by all protein crystallographers. However, macromolecular structures usually contain disordered or partially occupied fragments and other features not amenable to automatic and unequivocal interpretation by a ‘computer’. Static or dynamic disorder (see Glossary) of certain fragments is common in all protein models, even (or especially) of those refined at atomic resolution, as shown in Fig. 3. Despite the existence of sophisticated software capable of making this task easier, there is no better way than visual inspection of electron density maps to decide what is real and what is a spurious feature of a map. This task is accomplished using a graphics display and a software such as COOT [85]. A model must always be thoroughly scrutinized visually against electron density maps before accepting it as final.

There are very useful quality indicators that gauge the agreement between a model and the electron density map and that can be applied locally to selected model fragments, such as individual residues. They are calculated as the real-space correlation coefficient (RSCC) or real-space residual (RSR) or R-factor. The RSCC compares (on a suitable grid) the observed electron density calculated from the diffraction data, ρ_{obs}, with that calculated from the atomic model, ρ_{calc}: $\mathrm{RSCC} = [\Sigma|\rho_{obs} - <\rho_{obs}>|*\Sigma|\rho_{calc} + <\rho_{calc}>|]/[\Sigma|\rho_{obs} - <\rho_{obs}>|^2 * \Sigma|\rho_{calc} + <\rho_{calc}>|^2]^{1/2}$, where the summation runs over all grid points of a map encompassing a selected fragment of the model (e.g. an individual residue or ligand). The RSR corresponds to a

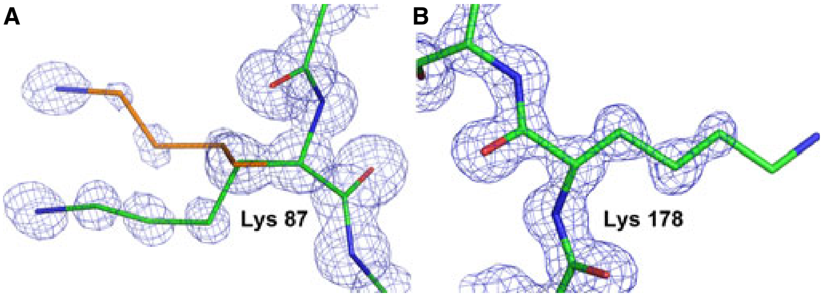

Fig. 3. Electron density for regions with disorder. (A) The model and the corresponding (F_{obs}, α_{calc}) map for statically disordered Lys87 in the structure of bovine trypsin (PDB code: 4I8G) [136], with its side chain in two conformations. The map was calculated at 0.8 Å resolution and displayed at the 1.4σ contour level. (B) Lys178 from *Erwinia chrysanthemi* L-asparaginase (PDB code: 1O7J) and the corresponding (F_{obs}, α_{calc}) map at 1.0 Å resolution, with well-defined main chain atoms but a dynamically disordered end of the side chain having no interpretable electron density.

normalized difference between the observed and calculated electron density, RSR = $\Sigma|\rho_{obs} - \rho_{calc}|/\Sigma|\rho_{obs} + \rho_{calc}|$, again calculated on a selected grid of points.

Model refinement

Structural models obtained from approximate experimentally-derived or MR-based phases are also only approximate and require further refinement. The refinement process usually involves alternating rounds of automated optimization of all refined parameters (e.g. according to least-squares or maximum-likelihood algorithms) and manual corrections of the model to improve its agreement with the electron density maps. These corrections are necessary because the automatically refined parameters may get stuck in a (mathematical) local minimum, instead of leading to the global, optimum solution. The model parameters that are optimized by refinement software include, for each atom, its x, y, z coordinates and a parameter reflecting its 'mobility' or smearing in space, known as the B-factor [or atomic displacement parameter (ADP), also referred to as the temperature factor]. B-factors are usually expressed in Å^2 and range from 2–100 Å^2. In theory, ADP should reflect the amplitude of atomic oscillations (which indeed increases with temperature) around an equilibrium position but, in reality, B is a capacious parameter, capable of absorbing many other effects, such as static disorder or even experimental errors. The B-factor model used is usually isotropic, at least for structures solved at a resolution of 1.4 Å or lower (i.e. it assumes that the displacements have the same amplitude in all directions). This model, represented by a sphere, is of course incorrect in an anisotropic environment but its huge advantage is simplicity and the introduction of only one parameter per atom. Structures refined at higher resolution often use the more correct anisotropic model of atomic displacements. Mathematical expression of the anisotropic B-factor, which is represented as a 3D ellipsoid, requires six parameters. This model is therefore very complex and 'data hungry' and must be used with caution to avoid overinterpreting the available experimental observations. As a compromise between isotropic and anisotropic models of vibrations of individual atoms, certain segments of a macromolecular structure (e.g. protein domains) can be treated as rigid bodies, assuming that they perform concerted motion. Such a motion is then described by anisotropic TLS parameters [86], where T stands for translational (linear) oscillations, L for librational (rotational) oscillations and S (screw) for the coupling of these two modes. TLS parametrization is quite popular in macromolecular refinement and can be aided by an algorithm (TLSMD) for macromolecule segmentation [87]. Selection of rigid groups should be reasonable and correspond to individual (sub)domains. An exceedingly large number of very small fragments unreasonably increases the number of refined parameters and again leads to models not fully justified by the experimental data. TLS groups should not be mixed with individual anisotropic B-factors. The group (TLS) anisotropic parameters are added to the individual atomic B_{iso} parameters. It should be noted that in PDB deposits, the TLS parameters are introduced twice, as matrices, and then duplicated in ANISOU records for each atom.

Even in the isotropic approximation, crystallographic models of macromolecules are tremendously complex. For example, a protein molecule of 20 kDa would take approximately 6000 parameters to refine. Frequently, the number of observations (i.e. measured unique reflections) (especially at low resolution, see below) is not quite sufficient. For this reason, the refinement is carried out under the control of stereochemical restraints that guide its progress by incorporating prior knowledge, or stereochemical common sense [88,89]. The most popular libraries of stereochemical restraints (their target values accompanied by standard uncertainties) were compiled based on small-molecule structures [90–92], although there is growing evidence from high-quality protein models that the nuances of macromolecular structures should also be taken into account [93]. For example, the N-Cα-C angle of the protein backbone has a wide distribution [94] and is correlated not only with residue type, but also with the local backbone conformation [95]. By modelling main-chain bond distances and angles as functions of backbone conformation, Tronrud and Karplus [96] were able to create a conformation-dependent stereochemical library, the use of which leads to better models. When stereochemical restraints are applied at medium or high resolution, they should not be enforced too tightly. Too tight restraints suppress the information from the physical experiment and lead to regularized rather than refined models. The rmsd of the final model from the targets should reflect the level of confidence of those targets. For bond distances, this value is approximately 0.01–0.02 Å.

The widely used refinement software REFMAC [97] includes many very useful options, such as iso/aniso/TLS parametrization, adjustment of stereochemical weights, automatic detection and refinement of crystal twinning, etc. PHENIX [98] offers a huge selection of adjustable options, conveniently hidden behind a graphics user interface. For refinement at atomic resolution (defined as at least 1.2 Å), the software of choice is SHELXL [99], which has evolved from the philosophy of small-molecule refinement to encompass macromolecules as well. In variance with REFMAC and PHENIX, which tend to be used for structure optimization against maximum-likelihood targets, SHELXL is based on the algorithm of least squares with accurate (in contrast to fast Fourier transform approximations) mathematical formulae. SHELXL is extremely flexible, allowing for practically any course of refinement. For example, it allows individual definitions of restraints, mixing of global/local parameters, refinement under target-less similarity restraints and restraining of ADPs.

Although many steps of crystal structure analysis have been automated, the interpretation of some fine features of electron density maps still requires a significant degree of human skill and experience [100]. A small degree of subjectivity is thus inevitable in this process and different individiuals working with the same data could occasionally produce slightly different results. It is, however, the role of statistical tests and validation criteria to determine whether such nuances would have any significance.

'One-click' structure solution and refinement

Many major improvements in structure determination and refinement, at least for the more routine cases, are a result of the availability of automated data collection facilities coupled with modern integrated software packages. The current trend in the structure determination process is to combine the best data reduction, structure determination (substructure solution, phasing, model building) and validation software and 'glue' them together with sophisticated and flexible structure determination protocols. The currently most commonly used pipelines are PHENIX [98,101], AUTO-RICKSHAW [102,103] and HKL-3000 [104]. The rapid progress in computer hardware, algorithms and the availability of multicore processors allows for almost 'real-time' structure determination. In some cases, structures can be determined even before all of the diffraction data have been collected. Such on-the-fly structure determination makes synchrotron trips, in contrast to remote access, very attractive. For example, experimenters working on structures of protein–ligand complexes, especially those that use soaks of cocktails of multiple ligands for screening [105], can obtain very rapid feedback regarding which ligands are bound and which are not, thus suggesting new soaking experiments that could be performed when still at the synchrotron. Automation can also significantly increase the throughput because the cocktail approach requires collecting of many data sets.

Structure determination for a crystal that has been soaked with a ligand (or a cocktail of ligands) is usually quite straightforward. For example, in HKL-3000 implementation, the data collection and reduction step is followed by electron density map generation using phases from the structure of the *apo*-protein, or derived by MR with MOLREP [62]. This is followed by sequential, semi-automated fit of each cocktail component into unexplained electron density. The analysis, performed by RESOLVE [83], produces a set of protein–ligand complex models, ranked by the quality of their fit to unassigned electron density. The recent software AUTODRUG performs these tasks fully automatically [106]. It should be noted that identification of the best fit often requires visual inspection of the ligand model because automated (and in a sense blind) ligand assignment can be impeded by conformational changes in the protein or partial disorder of the ligand structure. The top-ranked ligands are used to create new cocktails. Because rapid structure determination allows the data collection process to be more interactive, a protein's function (the holy grail of structural biology) might be elucidated according to this scheme during a single synchrotron trip.

In principle, the final refinement and validation of a newly-determined structure could also be performed at the synchrotron as this process is highly automated, especially for structures determined to a resolution higher than 2.5 Å. However, automatic refinement should not be trusted blindly and unconditionally, and it is strongly recommended that investigators inspect each structure visually before deposition. Parameters such as R, R_{free}, MOLPROBITY clash score, mean ADP (i.e. mean B-factor), etc., describe only the global quality of a structure. Visual inspection of the flexible parts of a structure, as identified by high atomic B-factors, can be important for proper interpretation of protein function.

Is the 'final' model really final?

Deposition of the atomic coordinates and structure factors, as well as publication of the relevant paper, used to be the final stages of the work of a crystallographer on a given problem, and these data could then be considered to be permanently stable. However, crystallographic procedures have been undergoing constant improvement; thus, structures that could be considered state-of-the-art at one time might not be completely satisfactory some decades later. For example, introduction of indicators such as R_{free} [107] (see Glossary) led to improvement in validation and prevention of overfitting in the refinement. In general, modern refinement techniques and better description of the models that lead to more realistic values of F_{calc} (inclusion of scattering from the bulk solvent and H atoms, TLS parameters, etc.), as well as F_{obs} (detection and correction of crystal twinning), result in improved description of macromolecular structures. Although the coordinates present in the PDB can be updated only by the original investigators (the periodic annotation changes instituted by the PDB should not, in principle, change the coordinates themselves), other crystallographers may sometimes be able to improve the deposited structures using the original data (that is one of the reasons why deposition of diffraction data is now mandatory). The practice of re-evaluating of the entire contents of the PDB and re-refining the structures in a consistent way has led to the development of PDB_REDO [108], a procedure and a site described as a 'constructive validation, more than just looking for errors'.

The pipeline utilized by PDB_REDO initially depended only on fully automatic refinement using REFMAC [97,109], without rebuilding by either manual or automated procedures. More recent implementations are capable of modifying the models by rebuilding the main chain and the side chains at problematic places and validating the results; if rebuilding improved the electron density maps, then the new coordinates would be retained. The procedure chooses the most appropriate B-factor model, tests different approaches to the definition of TLS groups (if any) and selects the most appropriate geometric restraint weights. However, even clear misassignment of metals is not corrected by PDB_REDO.

What to do with macromolecular structures from structural genomics centers?

Although most structures deposited in the PDB are accompanied by publications in scientific journals, that is not necessarily the case with structures determined by SG teams. This is not surprising because the stated aim of SG has been, until recently, to fill the gaps in the coverage of the protein fold space and to provide structural data for proteins from various sources that differed in their sequence by at least 70% from the proteins with known structure. Even the SG efforts directed against particular medically important targets (e.g. tuberculosis) have not led to full analysis and interpretation of the determined structures. According to the header records of the PDB files identified as output from SG, only 45% were accompanied by primary publications (unlike 86% of all structures in the PDB). In addition, refinement of the SG structures may not be as extensive and exhaustive as for other structures as a result of the pressure of time and the need to produce results rapidly. Nevertheless, SG structures are better on average than structures coming from traditional laboratories [110]. It is clear that the availability of SG-determined structures leaves open opportunities for their further refinement and/or deeper and more detailed analysis of their biological significance.

An example of such reinterpretation of a series of SG structures involved mapping the active site helix-to-strand conversion of CxxxxC peroxiredoxin Q enzymes [111]. In that rather unusual case, the RIKEN SG group provided Perkins *et al.* [111] with the original diffraction images that were then independently reprocessed. Reprocessing yielded significantly improved resolution (1.4 versus 1.6 Å, 2.0 versus 2.3 Å and 2.3 versus 2.6 Å for the three data sets in question), as well as improved scaling statistics. Utilization of better refinement protocols led to very significant decrease in the final *R* factors (e.g. *R* decreased from 20.0% to 12.0% and R_{free} decreased from 22.1% to 14.9% for the 1.4 Å structure) and it became possible to trace more residues, as well as water molecules. Most importantly, Perkins *et al.* [111] could provide a penetrating interpretation of the improved structures, adding significant value to the coordinates deposited in the PDB.

Are the ligands real?

A majority of the structures deposited in the PDB, and certainly all structures determined at a resolution higher than 2.5 Å, contain not only the coordinates of the atoms belonging to a protein or another macromolecule, but also coordinates of associated ligands and solvents. Such ligands are usually identified on the basis of electron density maps calculated after the completion of at least preliminary refinement of the protein coordinates. Because proteins are almost always crystallized from aqueous media, water molecules are present in the crystal lattice and many of them are ordered sufficiently well to be easily visible in the electron density maps. A comparatively small number of water molecules are buried deeply inside the protein, and such molecules have peaks of electron density that are as high as those for the surrounding macromolecule atoms. An example of such deeply buried water molecules is provided by bovine pancreatic trypsin inhibitor (BPTI), where four water molecules are seen in all structures of this small protein that contains only 58 amino acid residues [94,112,113]. Most water molecules, however, are located on the surface of the protein and their presence in the electron density is sometimes uncertain because of low occupancy and/or disorder.

How is it possible to be certain that a particular water molecule is real, as opposed to an artefact caused by noisy electron density? First of all, well-ordered water molecules must be hydrogen-bonded to polar atoms in the protein (if in the first hydration shell) or at least to some other validated water molecules (if in the second or third shell). Most well-ordered water molecules are involved, as either hydrogen bond donors or acceptors, in two to four such bonds. Both hydrogen bond distances and angles should be consistent with the tetrahedral pattern of hydrogen-bonding of water, although often there may be deviations from the typical values. As a guidance, the O…O distances in solid water (ice) are approximately 2.7 Å, no O…O hydrogen bonds can be shorter than 2.35 Å [114] and, consistent with van der Waals radii, the threshold values should be increased by 0.1 Å if O is replaced by an N atom. O…O distances > 3.2 Å should be disregarded as too long or representing very weak hydrogen bonds. On the other hand, the criteria should be implemented with 0.1–0.2 Å tolerance, which reflects the margin of error possible with (less accurate) macromolecular structures. Because four hydrogen bonds around a water molecule (assuming no bifurcation) are the theoretical maximum (two donors and two acceptors), the presence of more potential partners may indicate that the site is not occupied by a water molecule or that it is present in more than one orientation. A very good indication that the water molecules attached to the protein are real comes from their presence in the same location in multiple crystal structures, particularly in non-isomorphous crystals. In the BPTI example noted above, approximately half of the water molecules traced in two non-isomorphous structures determined at very high resolution (0.86–1.5 Å) were found in the same locations (within 1 Å after superposition), whereas the sites occupied by the remaining water molecules differed between the two structures. A similar situation is quite typical for other structures as well.

A large fraction of the protein structures contain identified ligands other than water molecules. Such ligands may be introduced into the crystals on purpose (e.g. as enzyme inhibitors co-crystallized with or soaked into their targets) or may have stayed bound throughout the purification of natural or recombinantly expressed proteins, or else may have been components of the crystallization/cryoprotection media. Stereochemical restraints for common ligands are present in standard libraries, although caution is advised in their use because errors are not infrequent [115]. More unique compounds may require prior knowledge of their structure and creation of appropriate restraints by the user.

Nonprotein covalent modifications that are quite common in the PDB structures are N-linked or (less commonly) O-linked carbohydrate molecules. Most proteins originating from eukaryotic organisms are glycosylated at the side chains of asparagine (in NXS/T sequences) or serine/threonine/tyrosine (no consensus sequence). However, such glycosylation patterns are often heterogeneous and the carbohydrates are well-ordered only if they are involved in intra- or intermolecular interactions [116]. Accordingly, it is usually not possible to trace the complete carbohydrate molecule and only one or a few sugar units closest to the site of attachment are visible. A somewhat similar situation is encountered with artificial affinity tags that are sometimes left attached to recombinant proteins submitted for crystallization. In most cases, such tags (non-native) are disordered and thus invisible in electron density maps but, in exceptional circumstances, they can become ordered or even control the crystal packing [117].

A script to assess the reliability of the published protein ligands (named *Twilight*) has been recently developed and used to evaluate all PDB structures deposited before 2012 [118]. The procedure utilizes the electron density maps that correspond to each PDB entry and are stored at the publicly accessible Uppsala Electron Density Server [119]. The script calculates the RSCCs for the electron density in the areas that correspond to the ligand atoms, flagging total correlations that are lower than 0.6, after correcting for the resolution of the analyzed structures. Problematic ligands were found in close to 3000 PDB entries, comprising just under 10% of all the coordinate sets that were evaluated. Unfortunately, one of the flagged structures originated in the laboratory of a coauthor of the present review, who is still not sure how an incorrect ligand could be fitted to such an excellent map (Fig. 4). Although many problems identified by *Twilight* were comparatively minor, some ligands that were very important for describing the enzymatic mechanisms were found to be either poorly fitted or not present at all. Alternative software that can accomplish such a ligand-validation task is VHELIBS [120].

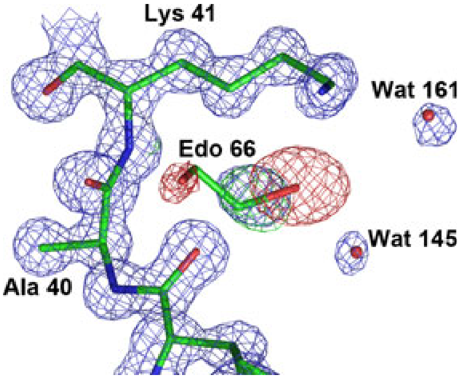

Fig. 4. An example of a phantom ligand in a protein structure refined at high resolution. Electron density and the atomic model are shown for a fragment of the cyclic form of BPTI refined at 1.0 Å resolution (PDB code: 1K6U). A weighted $2mF_{obs} - DF_{calc}$ map (blue) was contoured at 1.5σ, whereas the $mF_{obs} - DF_{calc}$ map was contoured at ± 2.5σ (green positive, red negative). It is clear that the ethylene glycol molecule was placed completely incorrectly and that, most likely, only a single water molecule should have been modelled in its place.

Misrepresentation of crystallographic experiments

Because crystal structure determination involves very sophisticated processing of experimental observations in many computational procedures, outright fabrication of crystallographic data is very infrequent, although several cases of faking diffraction data and/or the resulting structures have been uncovered. The first known case was a discovery that the published diffraction patterns attributed to valyl tRNA were actually those of human carbonic anhydrase B [121]. The substitution was detected by analyzing the unit cell parameters of the published diffraction photographs; their values are quite characteristic for a given crystal. Although such parameters might, by chance, bear similarity to those for crystals of other macromolecules, in this particular case, the latter possibility was ruled out through careful analysis of other aspects of the presented data.

In another case of scientific forgery, a number of structures of several different proteins, all originating from the same laboratory, had to be retracted after it had been reported that the data deposited in the PDB for the structure of protein C3b in the complement pathway (PDB code: 2HR0) were inconsistent with the known physical properties of macromolecular structures and their diffraction data [122]. The tell-tale signs of fabrication were the absence of the contribution of the bulk solvent to structure amplitudes at low-resolution, the fact that the electron density of the presumably largely unfolded domain was excellent, and the lack of correlation between surface accessibility and the atomic B-factors. In addition, some other features (18 distances between nonbonded atoms of less than 2 Å, several peptide torsion angles deviating from planarity by as much as 57° and 4.2% of outliers in the Ramachandran plot, almost all in one subunit) are clear indications of serious problems with this structure. That report incited further examination of other structures published by the same principal author and led to withdrawal from the PDB of a dozen coordinate sets, as well as to the retraction of the corresponding publications. The most highly cited among them was the structure of Dengue virus NS3 protease, published in 1999 (PDB code: 1BEF) and retracted in 2009. The unfortunate part of this story is that the particular publication which claimed to present a target for designing drugs against an important pathogen was cited almost 100 times in other scientific papers.

Even more recently, the structure of birch pollen hypoallergen Bet v 1d (PDB code: 3K78) was shown to have been fabricated [123]. One of the most damaging lines of evidence confirming that the deposited structure factors could not have been obtained from diffraction experiments was the fact that refinement of this comparatively low resolution (2.8 Å) structure against the deposited data, using a model with isotropic B-factors, resulted in R of 0.019 and R_{free} 0.040, values that were impossibly low. A number of other factors, such as unusual features of the electron density for residues with zero occupancy, provided additional evidence of the falsification of both the deposited structure and the 'experimental' data.

An encouraging aspect of these sad incidents is the demonstration that the community has the alertness and ability to detect fraudulent structures and eradicate them together with their perpetrators. In addition, these incidents have initiated campaigns of careful combing of the PDB deposits by several watchdog groups, a procedure that has an additional benefit of detecting other problems.

Honest errors in structure determination

Serious errors in describing a whole macromolecule are rare, especially nowadays, although errors in some local areas might be more common. A structure of ribulose-1,5-biphosphate carboxylase oxygenase with the chain of one of the subunits traced backwards was published [124], although the error was noted almost immediately [125]. A later re-enactment of this case [89] showed that, although it is possible to refine a backwards-traced structure at medium resolution to acceptable values of R and rmsd(bond), the value of R_{free} would remain completely unacceptable (in that case, 61.7%), clearly indicating that the model was in error. With the currently mandatory use of R_{free}, similar errors are unlikely to happen again.

A later case of an important series of structures that were seriously misinterpreted was due to deviation from established good standards in crystallographic procedures and to overinterpretation of low-resolution data. The structure of the MsbA ABC transporter protein [81], as well as several related structures published by the same group, had to be retracted after the structure of Sav1866, another member of the family, was published [126]. The structure of MsbA was refined by nonstandard protocols that utilized multiple molecular models. It must be emphasized that all these structures were very difficult to solve and even the apparently correct structure of Sav1866 is characterized by rather high values of R (25.5%) and R_{free} (27.2%), although such values are not unusual at 3 Å resolution.

Unlike the very rare cases mentioned above where the whole structures were questionable, local mistracing of elements of the protein chain has been more common, and several such cases were reviewed previously [89]. Although this type of error may have little importance if it happens to be limited to an area of the protein that is remote from the active site or from the site(s) of interaction with other proteins, in other cases, it may lead to misinterpretation of biological processes. One well-known case, where the modelling of a β-strand instead of a helix led to postulating a doubtful mechanism of autolysis, was provided by HIV-1 protease [127]. However, similar to the cases mentioned above, the implausibility of the original interpretation became clear almost immediately, when, first, the structure of a related RSV protease became available [128], and, soon thereafter, when the structure of HIV-1 protease itself was independently determined [129].

One of the important practical aspects of crystallographic structures is to provide details of the interactions between macromolecules (usually enzymes) and small- or large-molecule inhibitors. Interpretation of such structures depends very much on the quality of the electron density for the inhibitor. In some cases, such as the complex of botulinum neurotoxin type B protease with a small-molecule inhibitor BABIM [130], the structural conclusions had to be later retracted, although the crystallographic quality indicators appeared to be acceptable (resolution 2.8 Å, $R = 16.2\%$, $R_{\mathrm{free}} = 23.8\%$). Similarly, the validity of the structure of a complex of the same enzyme with a target peptide was questioned [131] because the 38-residue peptide was apparently fitted to a very noisy map that could not support the interpretation of its structure.

Structure validation using computer software and a crystallographer's brain

There are many lines of evidence to convince even the extreme skeptics that, when correctly determined, the crystallographic models provide a faithful representation of the biological structures. For example, multiple determinations (different polymorphs, different variants, multiple copies in the asymmetric unit) invariably show the same basic structural features; the macromolecule, even in a crystal, is surrounded by water (typically 50% by volume) and, if appropriately assayed, will carry out its biological function also within the crystal; molecular mechanisms deduced from crystal structures make logical sense and are also corroborated by solution studies, if these are available.

However, how can we be sure that a particular crystal structure determination provides *the best* representation of the macromolecule? To illustrate in a pervasive way how easy it is to accept structural results as correct just because they might lead to neat-looking figures, we introduced previously an 'enzyme', frankensteinase [1]. This 'protein' was put together from fragments of deposited structures, with the addition of a few improbable features produced by creative imagination. Many of the features shown in Fig. 5 emphasize the points made above, namely that, without critical assessment of the quality of the structures, reasonable agreement with previous knowledge, and without full understanding of the meaning of various numerical descriptors, it is possible to carry out the interpretation way beyond the limits of reality.

To gain maximum confidence in the crystallographic result, each structure should be thoroughly validated by being checked against standard criteria of crystallographic quality and all available a priori knowledge of chemical, stereochemical and biochemical properties appropriate for the investigated molecules. Sophisticated software is available that can be helpful for this purpose, which can be used to identify structural features that disagree with the accepted standards, typical for well-refined structures.

The most popular validation software is PROCHECK [132] and MOLPROBITY [133]. The former software is now somewhat obsolete and the latter one represents the currently recommended validation tool. Many validation procedures are built into graphics software (e.g. COOT)

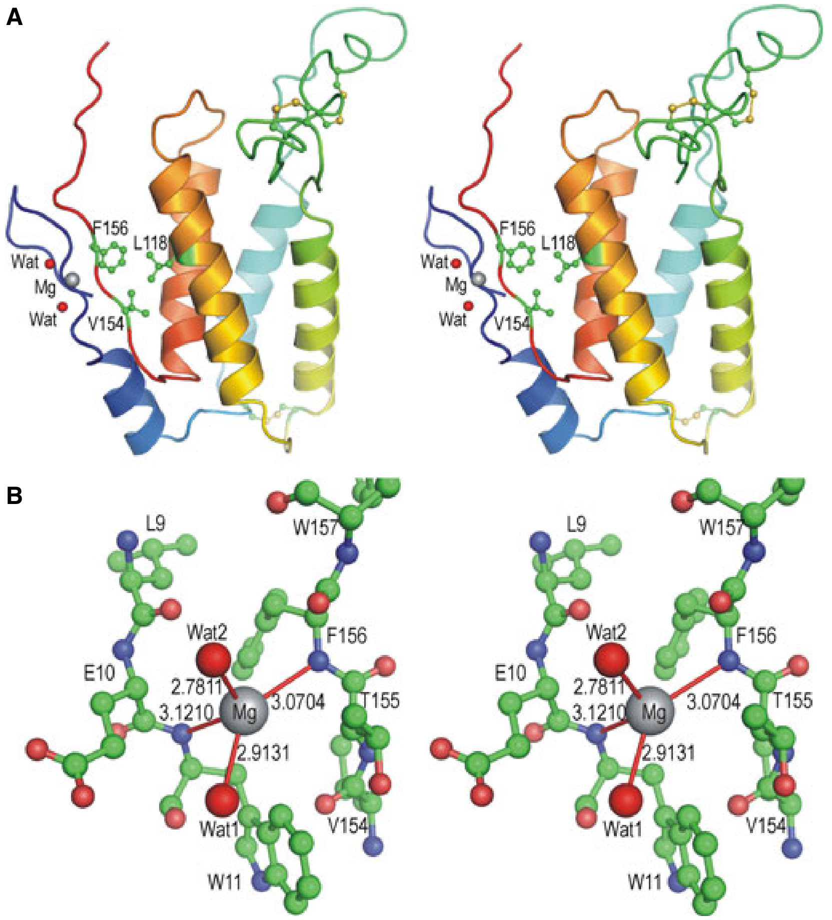

Fig. 5. Chain tracing and selected details of the 'enzyme' frankensteinase. A few problems with this structure need to be emphasized. (1) No such protein has ever existed, nor is likely to exist in the future. (2) The coordinates were freely taken from several real proteins but were assembled by the creators with a significant dose of imagination. (3) An 'active site' consisting of the hydrophobic side chains of phenylalanine, leucine and valine is rather unlikely to have catalytic properties. (4) Identification of a metal ion that is not properly coordinated by any part of the protein is rather doubtful. (5) The distances between the ion and the 'coordinating' atoms are shown with the precision of four decimal digits, vastly exceeding their accuracy. Besides, the 'bond' distances and 'coordination' by amide N-H groups are entirely unacceptable for magnesium. This figure was taken directly from our previously published review [1]. (A) A stereoview showing a tracing of the protein chain in rainbow colours, changing from the blue N terminus to red C terminus. Active site residues are in ball-and-stick rendering, the Mg^{2+} ion is shown as a grey ball, and water molecules as red spheres. (B) A detail of the Mg^{2+} binding site, with carbon atoms coloured in green, oxygen in red and nitrogen in blue.

[85]. Such software checks for a number of stereochemical properties of the whole structure and of individual residues. Among the validation criteria are the correctness of bond lengths and angles, the distribution of the main-chain torsion angles within the Ramachandran plot, the agreement of side-chain conformations with preferred rotamers, proper assignment of atoms in amide groups and histidine rings (HNQ), an absence of collisions between nonbonded atoms (steric clashes), planarity of aromatic rings and peptide groups, as well as a number of other indications. This type of validation can be used as an aid during model building and as the ultimate check before accepting the final model. Similar validation is performed by the PDB during the process of structure deposition.

The Ramachadran plot is often used to validate the correctness of a protein model because the values of the main-chain torsion angles are usually not restrained during refinement. Only selected combinations of the $\phi/\ \psi$ angles for residues other than glycine are allowed in a properly folded structure, although occasional deviations are observed (Fig. 6A). Such deviations must be strongly supported by the electron density. However, the Ramachandran plot for frankensteinase very clearly indicates that the model must be seriously wrong (Fig. 6B).

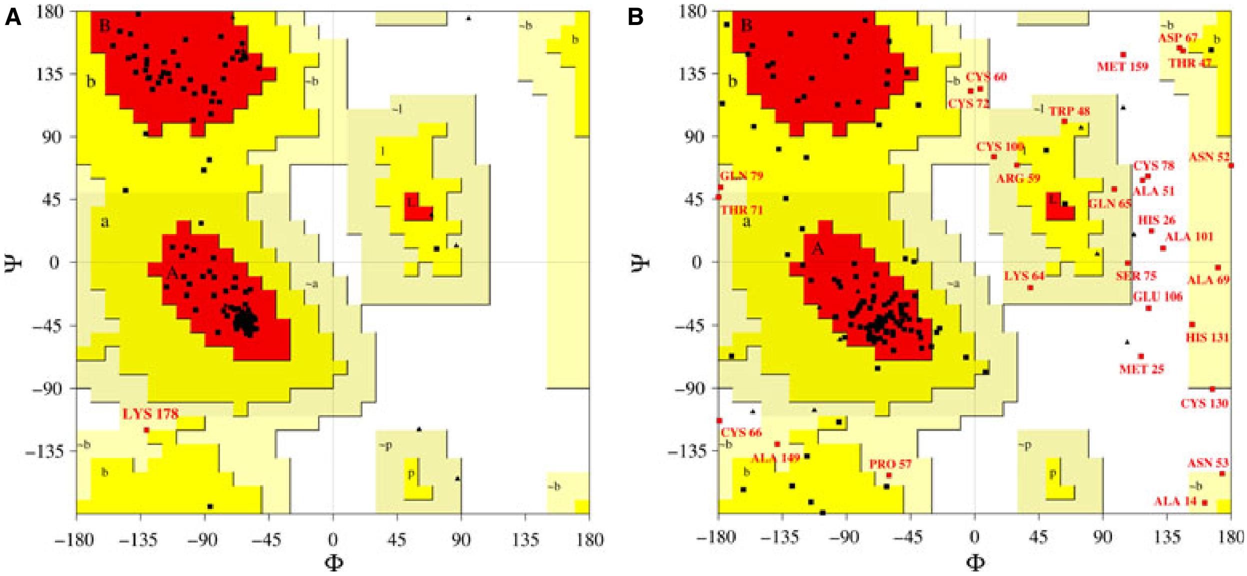

Fig. 6. Two examples of a Ramachandran plot. (A) A plot for *Erwinia chrysanthemi* L-asparaginase, one of the largest structures solved to date at atomic resolution (PDB code: 1O7J), where one of the lysine residues (Lys178; labelled) has an unusual main-chain conformation that is, however, strongly supported by the electron density shown in Fig. 3B. (B) A plot for the fictitious 'enzyme' frankensteinase characterized by a very large number of main-chain dihedral angle violations found outside of the allowed regions. Unfortunately, many of these outliers originated from a part of a protein taken from a legitimate PDB entry, which should remain anonymous.

Not all structural features can be validated automatically. Moreover, some features found in the crystal structures (especially at high resolution and of high quality) may be genuine but differ somewhat from the standards accepted by the validation software. Indeed, such places may carry the most important information about the function of a macromolecule. For example, strained conformations indicate an important role of a particular fragment or residue in the enzymatic mechanism or specific interactions between reacting molecules. It is therefore highly advisable to apply the additional validation tool in the form of a human brain. Only human knowledge allows making a decision if some 'abnormal' feature is spurious or if it indicates a genuinely interesting phenomenon of functional consequences.

Unfortunately, sometimes, this tool (the brain) is not applied properly or it is not equipped with the necessary chemical or biological knowledge. Chemical knowledge was evidently not applied to validate the PDB structure 3FJO, where 252 solvent sites are filled with Na^+ ions, forming a quite powerful charge bomb. A typical sign of inadequate application of scientific common sense are examples of 'numerology' or asinine 'cut-and-paste' activities, resulting in perpetuation of experimentally-derived values with implied unrealistically high precision. The publication accompanying the PDB structure 2C1C quotes cell dimensions as $a = 42.012$, $b = 58.3395$, $c = 146.5643$ Å, $\beta = 89.8739°$, although obviously such precision of cell parameters cannot be achieved from a diffraction experiment with a protein crystal.

It should be emphasized that a truly scientific activity always requires engagement of the human brain. Use of even the most sophisticated automatic procedures without a human thought is not sufficient.

Functional implications

There is one serious limitation in making conclusions based on structural biology experimental data in general, and crystallographic data in particular. The production of diffraction quality crystals is very difficult; quite often, proteins or other macromolecules are 'tortured' (genetically modified, exposed to highly unnatural chemical conditions, etc.) to promote crystallization. For example, the pH of the crystallization buffer may differ significantly from the pH that is optimal for functional studies. Similarly, the presence of artificial modifications to the protein's sequence to promote solubility or purification, such as polyhistidine affinity tags (His-tags) can unnaturally affect activity if, for example, the His-tag interacts with the active site. For these reasons, it must always be kept in mind that full characterization of biological macromolecules cannot be achieved by structural biology alone but requires a diverse set of tools and techniques.

The objective of determining the precise and accurate position of each and every atom in a macromolecular structure is an idealized goal in crystallography. As noted already, there are always fragments (at least solvent-exposed side chains) of increased mobility that will be reflected (at least) by elevated B-factors. In the worst-case scenario, there may be parts of a macromolecule that will be impossible to responsibly model because of lack of electron density as a result of total disorder. Although there is always the temptation to build as complete a model as possible, is not a good idea to use for this purpose suspicious, patchy and ephemeral electron density. The difficult question to face is then 'what to do with such regions?'. Unexpectedly, the best (optimistic of sorts) answer may be that we have just learned about a region that is genuinely disordered and that this observation may be of biological significance. Even if this difficult truth is accepted, there remains a practical issue of what to do with such fragments in PDB deposits. Some investigators try to patch such places with fictitious (but genuine in terms of sequence) residues and refine them freely, leading of course to sky-rocketing B-factors. Others refine such fragments with zero occupancy, meaning that they are insensitive to the experimental data. Both approaches are not to be recommended, although at least their outcome is easily detectable because the consumer of the PDB data takes the (highly advisable) additional effort of quickly scanning the atomic coordinate file. In another approach, investigators elect to model only those atoms that they think can be seen and to simply leave the rest out. Although this is quite honest, it creates interpretational problems of having residues of incorrect chemical composition. The very misleading remedy of replacing such faulty residues with chemically 'consistent' surrogates (usually Ala) is now largely phased out, although such structures are still encountered among the older deposits. In our opinion, the best approach is to omit such uninterpretable areas from the coordinate set altogether. There will be alerts (and occasionally problems) with the interpretation of such places but at least the consumers of those files will have a very clear warning and will not be invited to speculate about anything that has not been confirmed experimentally.

Another look at the PDB

Creation of the PDB was one of the major accomplishments of structural biology, especially making this unique resource of modern science freely available to the scientific community. The PDB, which has been in existence for just over 40 years, provides full and easy access to the results of decades of investigation of macromolecular structures by X-ray crystallography and other methods. By the time that the present review appears, it is likely that more than 100 000 structures will have been made available to the scientific community. Each (crystallographic) PDB deposit contains: (a) a header with information about the diffraction experiment, structure determination and refinement protocols used, additional flags (e.g. potential alerts) and other remarks (e.g. oligomerization, biological information, etc.); (b) the coordinates of atoms composing a structural model of the biological macromolecule; and, in a physically separate file, (c) a list of experimental structure factors that were derived from X-ray diffraction images, although older structures were often deposited without accompanying structure factors. In principle, the header should contain all of the information found in the Materials and methods section of a relevant publication. However, the information in the header is sometimes erroneous, incomplete, or both. [In the current PDB format, a value of 'NULL' for a given parameter indicates that the corresponding experimental value is missing.] The format of the PDB deposits which has been introduced over 40 years ago is limited (e.g. by the ability to include a maximum of 99 999 atoms, necessitating the splitting of some data sets) and will be changed to a more modern one soon. Useful computer graphics software exists to render atomic structural models in 3D for inspection and analysis together with the corresponding electron density maps (if the experimental structure factor amplitudes were also deposited). If necessary, the structures can therefore be rebuilt and re-refined. The PDB, being a repository of macromolecular structures, cannot intervene and modify the results provided by the original investigators, although it conducts validation of the submitted structures, giving the investigators a chance to correct possible errors and omissions. However, the ultimate responsibility for the veracity of the deposits rests exclusively with the depositors.

Several initiatives were recently started to further inspect and validate the correctness of all deposits (PDB_REDO) [108], their small molecule ligands (*Twilight*) [118] or metal ions [134]. These actions have led to the identification of several suboptimally determined structures and produced their improved models.

The overall quality of the deposited models and the ability of PDB users to examine or even re-interpret the original models makes protein crystallography a 'crown jewel' of experimental biomedical research. The contents of the PDB should not, however, be treated as the ultimate truth but, as always in science, as a starting point for further investigations.

Conclusions

The rather narrowly defined science of crystallography has an exceptionally central and interdisciplinary character among the natural and life sciences. It borders not only mathematics, physics, chemistry and biology, but also mineralogy and medicine. It has played, and continues to play, a key role in deciphering the 3D structure of all kinds of chemical molecules and, together with its variants, has been used to determine the structure of all kinds of biological macromolecules, from nucleic acids and proteins to viruses and the ribosome. It is also a 'friendly' science, appealing not only to the intellectual, but also to aesthetic faculties of even 'ordinary citizens', as illustrated by the UN-declared International Year of Crystallography (IYCr2014). Yet, for some inexplicable reason, it is no longer fashionable to teach crystallography to the next generation. Aware of this situation, we have prepared didactic reviews for this journal, first addressing novices and now the practicing but inexperienced protein crystallographers. In the present review, the goal has been to describe the basic principles as well as useful tricks-of-the-trade for the principal stages of protein crystal structure determination, which include crystallization, X-ray data collection, solution of the phase problem, and structure refinement and validation. Emphasis has been placed on how to avoid errors and pitfalls, as well as how to assess that macromolecular structures determined by others (deposited in the PDB and described in the literature) can be trusted and to what level of interpretation. In particular, readers will find advice about model validation methods, criteria and tools, including the most powerful tool, which is an educated, prepared and critical brain of the crystallographer him/herself. We have concluded the article with a rather extensive (albeit not exhaustive) glossary of important crystallographic terms and simple definitions. The present review will in no way replace a good handbook (which, fortunately, are in good supply), although it can serve as an introduction (and reference) for the more advanced texts.

For approximately 60 years, crystallography has been supplying accurate information about the 3D structure of the molecules of life; first, as a result of an extremely arduous and long process, and, today, with astonishingly powerful new methodology, at an ever increasing rate. The PDB will soon celebrate the accumulation of 100 000 macromolecular structure depositions. The crystal structures of macromolecules are usually of high quality, in the overwhelming part error-free and, as is demonstrated in numerous ways and examples, represent the true biological structures. As a community, we have to make sure that the high level of advancement is preserved and expanded, and also that the future generations of structural biologists are prepared to gain deeper insight from the massive amounts of data, and to take crystallography to areas that today may not even be foreseen.

Glossary of terms

Symmetry

Property of physical and mathematical objects. After a *symmetry operation*, a symmetrical object and its transformed copy are indistinguishable. *Proper* symmetry involves pure rotation. *Improper* symmetry combines rotation with reflection; in particular, improper *symmetry elements* include center of inversion ($\bar{1}$), mirror plane (m), and four-fold inversion axis ($\bar{4}$). Proper symmetry leaves the object unchanged, improper symmetry converts it into its mirror image. Chiral objects (e.g. native proteins or nucleic acids) are incompatible with improper symmetry. (However, a *pair* of enantiomers can have improper symmetry between them.) An important but trivial operation corresponds to 0° (or 360°) rotation (identity transformation). In classical crystallography, only the following rotations are compatible with the periodic nature of crystal lattices: one-, two-, three-, four- and six-fold. Symmetry transformations of finite objects (e.g. crystals) must leave at least one point stationary, and are governed by *point symmetry* elements. In infinite periodic crystal lattices, this requirement is not necessary, and symmetry transformations may include (a fraction of lattice-) translation. In particular, screw axes are possible. Some of them are right-handed (3_1, 4_1, 6_1, 6_2), some left-handed (3_2, 4_3, 6_4, 6_5) and some neutral (2_1, 4_2, 6_3).

Crystal system

All crystals are divided into seven groups, called crystal systems, according to their principal symmetry. From the lowest to the highest symmetry, the crystal systems are: *triclinic*, *monoclinic*, *orthorhombic*, *tetragonal*, *trigonal*, *hexagonal*, *cubic*. In another definition, the term 'crystal system' refers to the system of coordinates that most conveniently describes objects with a given symmetry. In general, those coordinate systems are non-Cartesian [i.e. can have axial units a, b, c, of any length and at any (not necessarily orthogonal) inclination α, β, γ, but must be compatible with the underlying symmetry].

Fractional and orthogonal coordinates

In crystallographic calculations (such as symmetry operations and Fourier summations), the atomic coordinates are expressed as fractions (*x, y, z*) of the unit cell edges with respect to the origin of the unit cell. However, for stereochemical calculations and graphics display, orthogonal coordinates (*X*, *Y*, *Z*), expressed in Å in a Cartesian system, are more convenient. In the PDB, the orthogonal atomic coordinates are stored. The transformation from the orthogonal to the fractional coordinates is trivial in the orthorhombic, tetragonal and cubic systems, where $x = X/a$, $y = Y/b$ and $z = Z/c$, although it is more complicated in the triclinic, monoclinic, trigonal and hexagonal systems.

Point group

A symmetrical object usually has more than one symmetry element. Those symmetry elements must form consistent sets, called groups. For finite objects, those groups are called point groups (or *crystal classes* with reference to crystals). The term 'group' is used in strict mathematical sense: it means that: (a) combination of any two symmetry elements gives another element of the group; (b) for each symmetry transformation, there is an inverse operation; and (c) the group includes a null (or identity) element. There are 32 3D crystallographic point groups but only 11 of them do not include improper symmetry. An international (Hermann–Mauguin) symbol of a point group lists the symmetry in the essential directions in each crystal system. For example, 2 is a monoclinic point group (two-fold axis in the b direction), 222 is an orthorhombic point group (three mutually perpendicular two-fold axes) and 432 is a cubic point group.

Lattice

A collection of nodes (i.e. points with integral coordinates along three basis vectors **a**, **b**, **c**). In crystallography, a lattice is an abstract representation of a crystal structure: it is periodic and infinite, and the real structure can be reconstructed by associating with each lattice point the concrete structural motif (molecule, cluster of ions or of molecules) that it symbolically represents. Strictly speaking, lattices with points at integral coordinates are called primitive (P) lattices. To preserve the maximal internal crystal symmetry, crystallography allows in some cases nodes with special combinations of 'half-integral' (i.e. ½) coordinates, resulting in the so-called centered Bravais lattices. All lattice points have exactly the same environment.

Reciprocal lattice

A mathematically abstract lattice based on vectors **a***, **b***, **c***, which have inverse-type relationship with the vectors of the direct (or real) lattice (e.g. $\mathbf{a}\cdot\mathbf{a}^* = 1$, $\mathbf{a}\cdot\mathbf{b}^* = 0$, etc.). These vectorial relations are very simple in orthogonal systems (e.g. $a^* = 1/a$, etc., for unit cell dimensions) but are rather complicated in general. Although theoretically abstract, the reciprocal lattice has a very practical interpretation as there is one-to-one correspondence between the diffraction pattern of a crystal and its reciprocal lattice. Thus, a reciprocal-lattice point with coordinates *hkl* corresponds exactly to Bragg reflection with indices *hkl*.

Unit cell

The smallest parallelepiped in the lattice whose translation (repetition) in the three lattice directions (vectors) **a**, **b**, **c** (which form its edges) recreates the entire crystal structure. From many possible choices, a conventional unit cell should be compatible with the symmetry of the space group. The smallest fragment, from which the entire unit cell can be recreated by symmetry, is called the *asymmetric unit*.

Bravais lattice

In some cases, nonprimitive unit cells have to be chosen to make the symmetry of the unit cell compatible with the symmetry of the entire lattice. Nonprimitive lattices, derived by Auguste Bravais, can have the following centering nodes: ½ ½ ½ (*I*), ½ ½ 0 (*C*) or 0 ½ ½, ½ 0 ½, ½ ½ 0 (*F*). Convention and symmetry considerations lead to 14 Bravais lattices in the seven crystal systems. The rhombohedral (*R*) unit cell represents a special case; it has nodes only at its vertices but has a three-fold axis along its body diagonal. Consequently, it has a unique shape with $a = b = c$ and $\alpha = \beta = \gamma$ (different than 90°).

Space group

Analogously to point groups, space groups are defined as consistent sets of symmetry elements of 3D lattices. A Hermann–Mauguin space-group symbol is formed by specifying the Bravais lattice and a list of symmetry elements in different directions, as in a point-group symbol. For example, $P2_12_12_1$ is an orthorhombic space group, with primitive lattice and two-fold screw axes parallel to **a**, **b** and **c**. There are 230 space groups, although only 65 of them do not include improper symmetry and are thus possible for macromolecular crystals.

Bragg's law

In Bragg's interpretation, the phenomenon of diffraction is viewed as reflection from the lattice planes (*hkl*). The incident and reflected beams make the same angle (θ) with the reflecting plane, and the two beams and the plane normal are coplanar, as in geometrical optics. The difference, however, is that this is *selective reflection*, which can occur only at θ angles selected according to the specific values of wavelength λ and interplanar spacing d_{hkl} by Bragg's law: $n\lambda = 2d_{hkl}\sin\theta$. This is because the rays reflected from consecutive planes must be in phase (i.e. must form an optical-path-difference equal to $n\lambda$). Note that the easy-to-measure angle between the incident and reflected beams is 2θ.

Friedel's law

In nonresonant situations (no special phenomena effected by the X-ray quanta in the electron clouds of atoms in a crystal), X-rays are reflected in the same fashion from both sides of a set of lattice planes. In effect, the diffraction pattern is centrosymmetric (identical reflection intensities on both sides of the origin). This is expressed by the equation $I(hkl) = I(\overline{hkl})$, known as Friedel's law. Friedel's law is violated in the presence of atoms that scatter a given wavelength anomalously.

Systematic absences

Translational symmetry causes extinctions among Bragg reflections, called systematic absences. Nonprimitive lattice centering wipes out all reflections with certain index categories. *I*-centering systematically extinguishes all reflections with $h + k + l$ odd, leaving only those for which $h + k + l = 2n$ (where n stands for any integer). The reflections present with *C*-centering are $h + k = 2n$, and with *F*-centering only those are left for which all three indices have the same parity. Screw axes extinguish reflections on axes running in the reciprocal-lattice direction corresponding to the direction of the screw, although the extinction rule depends on the order of the axis and its pitch (but it does not depend on the handedness of the screw). For example, a 2_1 screw along **b** affects reflection of the $0k0$ axis, leaving only those with $k = 2n$. A 6_1 (and 6_5) screw along **c** affects reflections of the $00l$ axis, leaving only those with $l = 6n$ (multiple of 6). The analogous rule for 6_2 (and 6_4) is $00l$ with $l = 3n$ and for 6_3 is $00l$ with $l = 2n$.

Structure factor

The physical quantity representing the amplitude and phase of the wave scattered by a crystal as reflection *hkl* is called the structure factor $F(hkl)$ and is calculated by adding up the contributions of all scattering atoms in the unit cell with a proper exponential (phase) factor accounting for the phase differences of the partial wavelets: $F(hkl) = \Sigma f_j \cdot \exp[2\pi i(hx_j + ky_j + lz_j)]$. Those phase (or optical-path) differences result from the spatial distribution of the scattering atoms. f_j is called the *atomic scattering factor* (or *formfactor*). It is obvious that, in general, $F = |F| \cdot \exp(i\alpha)$ is a complex quantity because it contains the imaginary unit i ($=\sqrt{-1}$). This is why it can express both the amplitude ('length' or modulus of F, $|F|$) and phase (direction or inclination angle of the vector F in the Argand diagram, α) of the scattered wave. The structure factor contains information about the atomic structure of the crystal because its calculation depends on the coordinates x_j, y_j, z_j of all atoms in the unit cell. Mathematically, the structure factor is the *Fourier transform* of the electron density in the crystal. The intensity of reflection *hkl* is proportional to the square of the amplitude: $I(hkl) = |F(hkl)|^2$ (scale neglected).

Fourier transform

In Fourier theory, a function defined as $F(h) = \Sigma f(x) \cdot \exp[2\pi i(hx)]$ has its almost identical twin companion, $f(x) = \Sigma F(h) \cdot \exp[-2\pi i(hx)]$ (scale neglected). In simple applications, these formulas (or Fourier transforms) can be interpreted as trigonometric Fourier series. The existence of this pair of Fourier transforms means that, if we have a recipe ($+i$ transform) for calculating F expressed as a Fourier series in f, then, automatically, f can be calculated as a Fourier series in F ($-i$ transform).

Phase problem

To be able to calculate electron density distribution in the crystal [$\rho(xyz)$] from the diffraction pattern using the Fourier transform $\rho(xyz) = (1/V)\Sigma \mathrm{F}(hkl) \cdot \exp[-2\pi i(hx + ky + lz)]$, there is a need to know the complete information (i.e. the magnitude and the phase) of each structure factor. However, because only reflection intensities, or squares of structure factor amplitudes, are measured experimentally, the information about the phases is not available. For this reason, the simple Fourier transform above cannot be used until the phase problem has been solved (i.e. until the phases have been obtained in one way or another).

Atomic scattering factor

The atomic scattering factor (or formfactor) f_j is the Fourier transform of the electron density (electron 'cloud') of a free atom of element *j*. For scattering of X-rays, it falls-off with the scattering angle, or $\sin\theta/\lambda$. At $\theta = 0$, *f* is equal to the atomic number (number of electrons). In normal (nonresonant) scattering of X-rays by electrons, *f* is real (*fo*). In resonant scattering, *f* becomes complex, which is expressed by the formula $fa = fo + f' + if'$, where *f'* is called the dispersive (real) and *if'* the anomalous (absorptive or imaginary) correction.

Atomic displacement parameters (ADPs)

Atoms in crystals are never motionless; instead, they oscillate around their equilibrium positions and, in addition, their positions may vary slightly in different crystal unit cells. Those effects (dynamic and static disorder) smear the electron density and make atomic scattering less efficient, especially at high angles. To account for this, the atomic scattering factor f_j must be multiplied by an exponential term with the ADP which takes the form $\exp[-B_j(\sin\theta/\lambda)^2]$. Sometimes, the parameter B is called the temperature factor. B is related to the displacement from the equilibrium position, *u*, in the following way: $B = 8\pi^2<u^2>$. Isotropic thermal model assumes the same (spherical) oscillation amplitude in all direction. This is not correct for bonded atoms, although the model is useful because it introduces only one parameter per atom (B_{iso}). The more adequate anisotropic model requires as many as six parameters per atom to define the general ellipsoid that describes the atomic motion. In protein models, it can only be used when the data resolution is higher than approximately 1.4 Å.

Electron density maps

Distribution of electrons, in the form of electron density ($e/\text{Å}^3$), usually drawn as a map, represents the chemical constituents of the crystal interior. Electron density maps are the primary product of crystal structure determination by X-ray crystallography, and atomic models represent their chemical interpretation. This is the case because the X-rays are scattered by electrons. In general, electron density is the Fourier transform of the structure factors, represented by reflection amplitudes and phases. It is calculated as a summation of contributions of all reflections over a grid of points within the unit cell of the crystal:

$$\rho(xyz) = 1/V\Sigma_{hkl}|A(hkl)|exp[\alpha(hkl)]exp[-2\pi i(hx+ky+lz)]$$

where $\rho(xyz)$ is the electron density at point *x*, *y*, *z*, $|A(hkl)|$ is the amplitude and $\alpha(hkl)$ is the phase of reflection *hkl* and the summation runs over all available reflections.

Several types of maps are used, depending on the kind of the amplitude |A| and phase α. At the first stages of structure solution, the observed (measured) amplitudes $|F_{\mathrm{obs}}|$ are used with phases estimated experimentally. For identification of the missing or incorrectly modelled structural features, the $F_{\mathrm{obs}} - F_{\mathrm{calc}}$ difference maps are used, calculated with differences between structure–factor amplitudes observed in the experiment and calculated from the current model, and with calculated phases:

$$\Delta\rho(xyz) = 1/V\Sigma_{hkl}(|F_{\mathrm{obs}}| - |F_{\mathrm{calc}}|)exp[\alpha_{\mathrm{calc}}]exp[-2\pi i(hx + ky + lz)]$$

Such a map shows positive density for new, unmodelled features and negative density for spurious fragments in the model. The use of $2F_{\mathrm{obs}} - F_{\mathrm{calc}}$ coefficients produces a map showing the electron density and the missing/spurious features simultaneously.

For the interpretation and rebuilding of the model during structure refinement, the electron density and difference maps are calculated using statistically σ_A weighted terms ($mF_{\mathrm{obs}} - DF_{\mathrm{calc}}$) and ($2mF_{\mathrm{obs}} - DF_{\mathrm{calc}}$), where the additional coefficients *m* and *D* take into account the imperfections of the current model and phases.

To obtain an unbiased representation of a problematic structural fragment, it is very useful to calculate an 'omit' map. Such a map is a difference map calculated after deleting the suspicious fragment (up to approximately 10%) from the structural model and refining the remaining model to remove phase bias. Such a map should reproduce the missing fragment without the effect of 'phase memory', which may persist in the phase set from the initial, wrongly interpreted stages of model building.

Patterson function

Before the phase problem is solved, the elegant Fourier transform based on the structure factors *F*(*hkl*) cannot be used to calculate the electron density map $\rho(xyz)$. However, when only the values of $|F(hkl)|^2$ are available, a similar Fourier transform can be calculated, called the Patterson map $P(uvw) = (1/V) \Sigma|F(hkl)|^2 \cdot \exp[-2\pi i(hu + kv + lw)]$. Mathematically, $P(uvw)$ is an autocorrelation function, or convolution of the structure with its centrosymmetric image. Therefore, although $\rho(xyz)$ represents the distribution of atoms, $P(uvw)$ represents the distribution of all possible interatomic vectors, all drawn from a common origin. For large structures (many atoms N), it contains a huge number of peaks (N^2) and is not amenable to straightforward interpretation, although it does contain information about the crystal structure and can help in its solution because each pair of atoms has a unique peak whose height is proportional to the product of the atomic numbers. The Patterson function is always centrosymmetric and contains (Harker) sections with accumulation of peaks corresponding to symmetry-related atoms.

Resolution

In principle, a faithful reconstruction of an image (crystal structure) would require the use of all (infinite number) $F(hkl)$ terms in the Fourier summation. This is impossible not only because of theoretical considerations (maximum θ or minimum d_{hkl} limitation in Bragg's law), but also for practical reasons because the 2θ angle has technical limitations and, especially, because protein crystals do not scatter X-rays to high angles. The minimum *d*-spacing corresponding to the highest θ angle at which measurable diffraction has been recorded, is known as the resolution of the diffraction pattern. The resolution in the reciprocal lattice has immediate interpretation in the direct space, corresponding to the ability to distinguish points ~*d* Å apart.

Residual or crystallographic *R*-factor

A measure of agreement of two (or several) sets of values, usually structure factor amplitudes $|F|$ or reflection intensities *I*. *R* is expressed as a fraction (often as a percentage) by calculating the sum of differences divided by the sum of all observations. For example, to monitor the refinement of a model, *R* (expected to be approximately 0.2 or lower for well-refined models) reports the agreement (or rather disagreement) between F_{calc} (calculated from the atomic coordinates) and F_{obs}: $R = (\Sigma||F_{\mathrm{obs}}| - |F_{\mathrm{calc}}||)/\Sigma|F_{\mathrm{obs}}|$. A better validation tool is R_{free}, calculated for a subset of approximately 1000 randomly selected reflections that have been excluded from any model refinement. A model 'improvement' that increases R_{free} is then immediately recognized as a false step (even if *R* drops). A typical 'sin' is the introduction of too many (unwarranted by the experimental data) model parameters, which then tend to replicate the errors of the data. If R_{free} is significantly higher (by > 0.08) than *R*, it signals overfitting (i.e. overinterpretation of the data).

In an analogous way, multiple observations of the same reflection intensities can be compared in R_{merge}. Sometimes *R* is used to compare real-space values (e.g. experimental and calculated electron density). In this variant, *R* can be used to pinpoint problematic areas of the model, whereas, as the reciprocal-space variant, it is a global indicator.

MIR and SIR

Isomorphous derivatives are crystal structures differing from the native one only by the presence of a few electron-rich (heavy) atoms. By comparing the diffraction patterns of the derivative and native crystals, the locations of the heavy markers can be determined and they become the source of phase information. With multiple derivatives (MIR), the phase problem can be solved unambiguously. With a single derivative (SIR), some extra information is necessary (e.g. from anomalous scattering) to resolve the ambiguity.

SAD/MAD

By tuning the energy of X-ray photons to resonance with special (anomalous) atoms in the structure, the anomalous signal responsible for the breakdown of Friedel's law can be enhanced. By measuring complete data sets at several wavelengths (MAD) near the absorption (resonance) edge (e.g. at the inflection point of the edge, at the absorption peak and at its high-energy tail), the anomalous scatterers can be located and then the reflection phases can be calculated analytically. The advantage of this method is the use of only one crystal with a suitable anomalous scatterer, usually selenium introduced into the protein sequence recombinantly in the form of Se-Met (as a substitute for Met). In the SAD variant, the simplification goes even farther because only one data set with the anomalous signal is collected.

MR

A method for the solution of the phase problem based on the Patterson function and an existing approximate search model for the unknown structure. The atomic model provides a set of interatomic vectors, which are then matched against the peaks of the Patterson function of the unknown structure. If the model is sufficiently similar, the search algorithm will detect its orientation and translation in the unit cell.

Direct methods

Classic direct methods (DM) of phase determination exploit mathematical relationships (equalities and inequalities) between structure factors that restrict or fix their phases. The theory of DM is based on statistical laws of structure factor distributions, and so in essence the relationships are probabilistic. For the success of DM, the data must have atomic (i.e. 1.2 Å or higher) resolution and the number of (non-H) atoms to locate (N) cannot be too large (maximum of approximately 1000) because the phase probabilities have the square-root of N in the denominator. DM can be quite successful even at low resolution, however, when the goal is to locate a substructure of heavy atoms, where the distances between atoms are typically much longer than 1.2 Å. A variant of DM, which, in the strict sense, work in the reciprocal space, are *ab initio* methods, such as shake-and-bake, which use a dual-space approach to phase solution. Here, the phases are obtained by iterative application of the phase determination formulas in the reciprocal space and of educated discrimination between the potential atomic peaks in the electron density maps calculated inbetween.

Ramachandran plot

Because of the simple and repetitive nature of the atomic groups forming the polypeptide chain (…-N-Cα-CO-…), there are also only three torsion angles that are repeated along a polypeptide chain: ϕ (CO-N-Cα-CO), ψ (N-Cα-CO-N) and ω (Cα-N-CO-Cα). The ω peptide bond torsion angle is usually close to 180° (for *trans* peptides) or 0° (for the rare *cis* peptides) but the ϕ/ψ angles for each residue are variable. Ramachandran showed that almost all pairs of ϕ/ψ values are forbidden on a conformational map because of atomic collisions. The only allowed regions are for ϕ/ψ angles of approximately –60°/–60° and –120°/120°. When repeated, the former combination leads to an α-helix, and the latter to a chain in an extended β-conformation. Today, the Ramachandran plot is used in the opposite sense, namely to verify the correctness of the conformation of an experimental (or sometimes theoretical) model of a polypeptide chain. The original Ramachandran map is displayed as an energy-contour plot in the coordinates of the ϕ/ψ angles, and the actual ϕ/ψ values are marked against this background. A correctly folded polypeptide chain should have > 90% of all residues in the most favoured regions of the Ramachandran plot.

Disorder

Disordered fragments are those fragments of the crystal structure that are perturbed from the ideal periodicity of the crystal lattice and therefore do not follow the perfect crystalline ‘order’. Static disorder occurs when some atoms are located in somewhat different positions in different unit cells of the crystal. Dynamic disorder is related to (thermal) vibrations of atoms around their equilibrium position that smear them during the X-ray exposure. Both effects result in smearing of the electron density proportionally to the degree of disorder, up to the totally featureless level in the region of the bulk solvent. A small degree of disorder is usually described by the ADPs. Static disorder can often be modelled by multiple (rarely more than two) conformations of the relevant fragment with fractional occupancies. Severely disordered fragments are very difficult or simply impossible to model.

Acknowledgements

We thank Dr Jerry Alexandratos for assistance with the preparation of the figures and Milosz Ruszkowski for the photograph in Fig. 1A. Original work in the laboratories of A.W. and Z.D. was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. W.M. was supported by grants GM053163, GM094585, GM094662 and GM093342, and was funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272201200026C.

References

1 Wlodawer A, Minor W, Dauter Z & Jaskolski M (2008) Protein crystallography for non-crystallographers or how to get the best (but not more) from the published macromolecular structures. *FEBS J* **275**, 1–21.

2 Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rogers JR, Kennard O, Shimanouchi T & Tasumi M (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* **112**, 535–547.

3 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN & Bourne PE

(2000) The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242.

4 Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H & Phillips DC (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181**, 662–666.

5 Levitt M (2007) Growth of novel protein structural data. *Proc Natl Acad Sci USA* **104**, 3183–3188.

6 Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G & North ACT (1960) Structure of haemoglobin. A three-dimensional fourier synthesis at 5.5-Å resolution, obtained by X-ray analysis. *Nature* **185**, 416–421.

7 Wlodawer A, Hodgson KO & Shooter EM (1975) Crystallization of nerve growth factor from mouse submaxillary glands. *Proc Natl Acad Sci USA* **72**, 777–779.

8 McDonald NQ, Lapato R, Murray-Rust J, Gunning J, Wlodawer A & Blundell TL (1991) New protein fold revealed by a 2.3 Å resolution crystal structure of NGF. *Nature* **354**, 411–414.

9 Barbieri M, Pettazzoni P, Bersani F & Maraldi NM (1970) Isolation of ribosome microcrystals. *J Mol Biol* **54**, 121–124.

10 Wittmann HG, Mussig J, Piefke J, Gewitz HS, Rheinberger HJ & Yonath A (1982) Crystallization of *Escherichia coli* ribosomes. *FEBS Lett* **146**, 217–220.

11 Ban N, Nissen P, Hansen J, Moore PB & Steitz TA (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**, 905–920.

12 Wimberly BT, Brodersen DE, Clemons WM Jr, Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T & Ramakrishnan V (2000) Structure of the 30S ribosomal subunit. *Nature* **407**, 327–339.

13 Schluenzen F, Tocilj A, Zarivach R, Harms J, Gluehmann M, Janell D, Bashan A, Bartels H, Agmon I, Franceschi F *et al.* (2000) Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. *Cell* **102**, 615–623.

14 Rupp B (2009) Biomolecular Crystallography Principles, Practice, and Applicaton to Structural Biology. Garland Science Titles, New York.

15 Drenth J (1999) Principles of Protein X-ray Crystallography. Springer-Verlag, New York.

16 McPherson A (2003) Introduction to Macromolecular Crystallography. Wiley-Liss, Inc., New York.

17 Blow D (2002) Outline of Crystallography for Biologists. Oxford University Press, New York.

18 Rhodes G (2006) Crystallography Made Crystal Clear. Academic Press, Burlington.

19 Blundell TL & Johnson LN (1976) Protein Crystallography. Academic Press, New York.

20 Hahn T (ed) (2005) International Tables for X-Ray Crystallography vol A, 5th edn. Springer, New York.

21 Dauter Z & Jaskólski M (2010) How to read (and understand) Volume A of *International Tables for Crystallography*: an introduction for nonspecialists. *J Appl Cryst* **43**, 1150–1171.

22 Minor DL Jr (2007) The neurobiologist's guide to structural biology: a primer on why macromolecular structure matters and how to evaluate structural data. *Neuron* **54**, 511–533.

23 Brown EN & Ramaswamy S (2007) Quality of protein crystal structures. *Acta Crystallogr* **D63**, 941–950.

24 Bagaria A, Jaravine V & Guntert P (2013) Estimating structure quality trends in the Protein Data Bank by equivalent resolution. *Comput Biol Chem* **46C**, 8–15.

25 Reichert ET & Brown AP (1909) The Differentiation and Specificity of Corresponding Proteins and Their Vital Substances in Relation to Biological Classification and Organic Evolution: The Crystallography of Hemoglobins. Carnegie Institution of Washington, Washington, DC.

26 McPherson A (1999) Crystallization of Biological Macromolecules. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

27 Ducruix A & Giegé R (1999) Crystallization of Nucleic Acids and Proteins: A Practical Approach, 2nd edn. Oxford University Press, New York.

28 Bergfors T (2009) Protein Crystallization, 2nd edn. International University Line, La Jolla, CA USA.

29 Bukowska MA & Grutter MG (2013) New concepts and aids to facilitate crystallization. *Curr Opin Struct Biol* **23**, 409–416.

30 Deisenhofer J, Epp O, Miki K, Huber R & Michel H (1984) X-ray structure analysis of a membrane protein complex. Electron density map at 3 Å resolution and a model of the chromophores of the photosynthetic reaction center from *Rhodopseudomonas viridis*. *J Mol Biol* **180**, 385–398.

31 Jiang Y, Lee A, Chen J, Cadene M, Chait BT & MacKinnon R (2002) Crystal structure and mechanism of a calcium-gated potassium channel. *Nature* **417**, 515–522.

32 Jiang Y, Lee A, Chen J, Ruta V, Cadene M, Chait BT & MacKinnon R (2003) X-ray structure of a voltage-dependent K^+ channel. *Nature* **423**, 33–41.

33 Rasmussen SG, Choi HJ, Rosenbaum DM, Kobilka TS, Thian FS, Edwards PC, Burghammer M, Ratnala VR, Sanishvili R, Fischetti RF *et al.* (2007) Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. *Nature* **450**, 383–387.

34 Day PW, Rasmussen SG, Parnot C, Fung JJ, Masood A, Kobilka TS, Yao XJ, Choi HJ, Weis WI, Rohrer DK *et al.* (2007) A monoclonal antibody for G protein-coupled receptor crystallography. *Nat Methods* **4**, 927–929.

35 Steyaert J & Kobilka BK (2011) Nanobody stabilization of G protein-coupled receptor conformational states. *Curr Opin Struct Biol* **21**, 567–572.

36 Zou Y, Weis WI & Kobilka BK (2012) N-terminal T4 lysozyme fusion facilitates crystallization of a G protein coupled receptor. *PLoS ONE* **7**, e46039.

37 Niimura N, Arai S, Kurihara K, Chatake T, Tanaka I & Bau R (2006) Recent results on hydrogen and hydration in biology studied by neutron macromolecular crystallography. *Cell Mol Life Sci* **63**, 285–300.

38 Coates L, Tuan HF, Tomanicek S, Kovalevsky A, Mustyakimov M, Erskine P & Cooper J (2008) The catalytic mechanism of an aspartic proteinase explored with neutron and X-ray diffraction. *J Am Chem Soc* **130**, 7235–7237.

39 Phillips JC, Wlodawer A, Yevitz MM & Hodgson KO (1976) Applications of synchrotron radiation to protein crystallography: preliminary results. *Proc Natl Acad Sci USA* **73**, 128–132.

40 Dauter Z, Jaskolski M & Wlodawer A (2010) Impact of synchrotron radiation on macromolecular crystallography: a personal view. *J Synchrotron Radiat* **17**, 433–444.

41 Boutet S, Lomb L, Williams GJ, Barends TR, Aquila A, Doak RB, Weierstall U, DePonte DP, Steinbrener J, Shoeman RL *et al.* (2012) High-resolution protein structure determination by serial femtosecond crystallography. *Science* **337**, 362–364.

42 Barends TR, Foucar L, Shoeman RL, Bari S, Epp SW, Hartmann R, Hauser G, Huth M, Kieser C, Lomb L *et al.* (2013) Anomalous signal from S atoms in protein crystallographic data from an X-ray free-electron laser. *Acta Crystallogr* **D69**, 838–842.

43 Diederichs K, McSweeney S & Ravelli RB (2003) Zero-dose extrapolation as part of macromolecular synchrotron data reduction. *Acta Crystallogr* **D59**, 903–909.

44 Diederichs K (2006) Some aspects of quantitative analysis and correction of radiation damage. *Acta Crystallogr* **D62**, 96–101.

45 Popov AN & Bourenkov GP (2003) Choice of data-collection parameters based on statistic modelling. *Acta Crystallogr* **D59**, 1145–1153.

46 Bourenkov GP & Popov AN (2006) A quantitative approach to data-collection strategies. *Acta Crystallogr* **D62**, 58–64.

47 Mueller M, Wang M & Schulze-Briese C (2012) Optimal fine ϕ-slicing for single-photon-counting pixel detectors. *Acta Crystallogr* **D68**, 42–56.

48 Weiss MS & Hilgenfeld R (1997) On the use of the merging *R* factor as a quality indicator for X-ray data. *J Appl Cryst* **30**, 203–205.

49 Diederichs K & Karplus PA (1997) Improved *R*-factors for diffraction data analysis in macromolecular crystallography. *Nat Struct Biol* **4**, 269–275.

50 Weiss MS (2001) Global indicators of X-ray data quality. *J Appl Cryst* **34**, 130–135.

51 Evans PR (2011) An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallogr* **D67**, 282–292.

52 Karplus PA & Diederichs K (2012) Linking crystallographic model and data quality. *Science* **336**, 1030–1033.

53 Evans PR & Murshudov GN (2013) How good are my data and what is the resolution? *Acta Crystallogr* **D69**, 1204–1214.

54 Schneider TR & Sheldrick GM (2002) Substructure solution with SHELXD. *Acta Crystallogr* **D58**, 1772–1779.

55 Dauter Z, Botos I, LaRonde-LeBlanc N & Wlodawer A (2005) Pathological crystallography: case studies of several unusual macromolecular crystals. *Acta Crystallogr* **D61**, 967–975.

56 Yeates TO (1997) Detecting and overcoming crystal twinning. *Methods Enzymol* **276**, 344–358.

57 Sheldrick GM (1990) Phase annealing in SHELX-90: direct methods for larger structures. *Acta Crystallogr* **A46**, 467–473.

58 Sheldrick GM (2002) Macromolecular phasing with SHELXE. *Z. Kristallogr* **217**, 644–650.

59 Morris RJ & Bricogne G (2003) Sheldrick's 1.2 Å rule and beyond. *Acta Crystallogr* **D59**, 615–617.

60 Thaimattam R, Tykarska E, Bierzynski A, Sheldrick GM & Jaskolski M (2002) Atomic resolution structure of squash trypsin inhibitor: unexpected metal coordination. *Acta Crystallogr* **D58**, 1448–1461.

61 McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC & Read RJ (2007) *Phaser* crystallograhic software. *Appl Cryst* **40**, 658–674.

62 Vagin A & Teplyakov A (1997) MOLREP: an automated program for molecular replacement. *J Appl Cryst* **30**, 1022–1025.

63 DiMaio F, Terwilliger TC, Read RJ, Wlodawer A, Oberdorfer G, Wagner U, Valkov E, Alon A, Fass D, Axelrod HL *et al.* (2011) Improved molecular replacement by density and energy guided protein structure optimization. *Nature* **473**, 540–543.

64 Li M, DiMaio F, Zhou D, Gustchina A, Lubkowski J, Dauter Z, Baker D & Wlodawer A (2011) Crystal structure of XMRV protease differs from the structures of other retropepsins. *Nat Struct Mol Biol* **18**, 227–229.

65 Khatib F, DiMaio F, Cooper S, Kazmierczyk M, Gilski M, Krzywda S, Zabranska H, Pichova I, Thompson J, Popovic Z *et al.* (2011) Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nat Struct Mol Biol* **18**, 1175–1177.

66 Gilski M, Kazmierczyk M, Krzywda S, Zabranska H, Cooper S, Popovic Z, Khatib F, DiMaio F, Thompson J, Baker D *et al.* (2011) High-resolution structure of a retroviral protease folded as a monomer. *Acta Crystallogr* **D67**, 907–914.

67 Long F, Vagin AA, Young P & Murshudov GN (2008) BALBES: a molecular-replacement pipeline. *Acta Crystallogr* **D64**, 125–132.

68 Keegan RM & Winn MD (2008) MrBUMP: an automated pipeline for molecular replacement. *Acta Crystallogr* **D64**, 119–124.

69 Hendrickson WA, Horton JR & LeMaster DM (1990) Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three dimensional structure. *EMBO J* **9**, 1665–1672.

70 Dauter Z, Dauter M & Rajashankar KR (2000) Novel approach to phasing proteins: derivatization by short cryo soaking with halides. *Acta Crystallogr* **D56**, 232–237.

71 Hendrickson WA & Teeter MM (1981) Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulfur. *Nature* **290**, 107–113.

72 Dauter Z, Dauter M, de la Fortelle E, Bricogne G & Sheldrick GM (1999) Can anomalous signal of sulfur become a tool for solving protein crystal structures? *J Mol Biol* **289**, 83–92.

73 Dauter Z & Adamiak DA (2001) Anomalous signal of phosphorus used for phasing DNA oligomer: importance of data redundancy. *Acta Crystallogr* **D57**, 990–995.

74 Liu Q, Dahmane T, Zhang Z, Assur Z, Brasch J, Shapiro L, Mancia F & Hendrickson WA (2012) Structures from anomalous diffraction of native biological macromolecules. *Science* **336**, 1033–1037.

75 Liu Q, Liu Q & Hendrickson WA (2013) Robust structural analysis of native biological macromolecules from multi-crystal anomalous diffraction data. *Acta Crystallogr* **D69**, 1314–1332.

76 Dauter Z, Dauter M & Dodson E (2002) Jolly SAD. *Acta Crystallogr* **D58**, 494–506.

77 Hendrickson WA (1991) Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254**, 51–58.

78 Hendrickson WA (1999) Maturation of MAD phasing for the determination of macromolecular structures. *J Synchr Rad* **6**, 845–851.

79 Burnley BT, Afonine PV, Adams PD & Gros P (2012) Modelling dynamics in protein crystal structures by ensemble refinement. *Elife* **1**, e00311.

80 Fraser JS, Clarkson MW, Degnan SC, Erion R, Kern D & Alber T (2009) Hidden alternative structures of proline isomerase essential for catalysis. *Nature* **462**, 669–673.

81 Chang G & Roth CB (2001) Structure of MsbA from *E. coli*: a homolog of the multidrug resistance ATP binding cassette (ABC) transporters. *Science* **293**, 1793–1800.

82 Perrakis A, Morris R & Lamzin VS (1999) Automated protein model building combined with iterative structure refinement. *Nat Struct Biol* **6**, 458–463.

83 Terwilliger TC (2003) SOLVE and RESOLVE: automated structure solution and density modification. *Methods Enzymol* **374**, 22–37.

84 Cowtan K (2006) The *Buccaneer* software for automated model building. 1. Tracing protein chains. *Acta Crystallogr* **D62**, 1002–1011.

85 Emsley P & Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr* **D60**, 2126–2132.

86 Busing WR & Levy HA (1964) The effect of thermal motion on the estimation of bond lengths from diffraction measurements. *Acta Crystallogr* **17**, 142–146.

87 Painter J & Merritt EA (2006) Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr* **D62**, 439–450.

88 Hendrickson WA (1985) Stereochemically restrained refinement of macromolecular structures. *Methods Enzymol* **115**, 252–270.

89 Kleywegt GJ & Jones TA (1995) Where freedom is given, liberties are taken. *Structure* **3**, 535–540.

90 Engh R & Huber R (1991) Accurate bond and angle parameters for X-ray protein-structure refinement. *Acta Crystallogr* **A47**, 392–400.

91 Engh RA & Huber R (2001) Structure quality and target parameters. In *International Tables for Crystallography*, (M.G. Rossmann & E. Arnold, eds) pp. 382–392. Kluwer Academic Publishers, Dordrecht.

92 Allen FH (2002) The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr* **B58**, 380–388.

93 Jaskolski M, Gilski M, Dauter Z & Wlodawer A (2007) Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? *Acta Crystallogr* **D63**, 611–620.

94 Addlagatta A, Krzywda S, Czapinska H, Otlewski J & Jaskolski M (2001) Ultrahigh-resolution structure of a BPTI mutant. *Acta Crystallogr* **D57**, 649–663.

95 Berkholz DS, Shapovalov MV, Dunbrack RL Jr & Karplus PA (2009) Conformation dependence of backbone geometry in proteins. *Structure* **17**, 1316–1325.

96 Tronrud DE & Karplus PA (2011) A conformation-dependent stereochemical library improves crystallographic refinement even at atomic resolution. *Acta Crystallogr* **D67**, 699–706.

97 Murshudov GN, Skubak P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, Winn MD, Long F & Vagin AA (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr* **D67**, 355–367.

98 Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr* **D66**, 213–221.

99 Sheldrick GM (2008) A short history of SHELX. *Acta Crystallogr* **A64**, 112–122.

100 Kleywegt GJ (2000) Validation of protein crystal structures. *Acta Crystallogr* **D56**, 249–265.

101 Adams PD, Grosse-Kunstleve RW, Hung LW, Ioerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK & Terwilliger TC (2002) PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr* **D58**, 1948–1954.

102 Panjikar S, Parthasarathy V, Lamzin VS, Weiss MS & Tucker PA (2005) Auto-rickshaw: an automated crystal structure determination platform as an efficient tool for the validation of an X-ray diffraction experiment. *Acta Crystallogr* **D61**, 449–457.

103 Panjikar S, Parthasarathy V, Lamzin VS, Weiss MS & Tucker PA (2009) On the combination of molecular replacement and single-wavelength anomalous diffraction phasing for automated structure determination. *Acta Crystallogr* **D65**, 1089–1097.

104 Minor W, Cymborowski M, Otwinowski Z & Chruszcz M (2006) HKL-3000: the integration of data reduction and structure solution – from diffraction images to an initial model in minutes. *Acta Crystallogr* **D62**, 859–866.

105 Shumilin IA, Cymborowski M, Chertihin O, Jha KN, Herr JC, Lesley SA, Joachimiak A & Minor W (2012) Identification of unknown protein function using metabolite cocktail screening. *Structure* **20**, 1715–1725.

106 Tsai Y, McPhillips SE, Gonzalez A, McPhillips TM, Zinn D, Cohen AE, Feese MD, Bushnell D, Tiefenbrunn T, Stout CD *et al.* (2013) AutoDrug: fully automated macromolecular crystallography workflows for fragment-based drug discovery. *Acta Crystallogr* **D69**, 796–803.

107 Brünger AT (1992) The free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–474.

108 Joosten RP, Joosten K, Murshudov GN & Perrakis A (2012) PDB_REDO: constructive validation, more than just looking for errors. *Acta Crystallogr* **D68**, 484–496.

109 Murshudov GN, Vagin AA & Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr* **D53**, 240–255.

110 Chruszcz M, Domagalski M, Osinski T, Wlodawer A & Minor W (2010) Unmet challenges of structural genomics. *Curr Opin Struct Biol* **20**, 587–597.

111 Perkins A, Gretes MC, Nelson KJ, Poole LB & Karplus PA (2012) Mapping the active site helix-to-strand conversion of CxxxxC peroxiredoxin Q enzymes. *Biochemistry* **51**, 7638–7650.

112 Wlodawer A, Deisenhofer J & Huber R (1987) Comparison of two highly refined structures of bovine pancreatic trypsin inhibitor. *J Mol Biol* **193**, 145–156.

113 Czapinska H, Otlewski J, Krzywda S, Sheldrick GM & Jaskolski M (2000) High-resolution structure of bovine pancreatic trypsin inhibitor with altered binding loop sequence. *J Mol Biol* **295**, 1237–1249.

114 Jaskolski M (1978) Very short hydrogen bond: X-ray structure of 2.6-dimethylpyridine N-oxide semiperchlorate. *Pol J Chem* **52**, 2399–2404.

115 Jaskolski M (2013) On the propagation of errors. *Acta Crystallogr* **D69**, In press.

116 Pietrzyk AJ, Panjikar S, Bujacz A, Mueller-Dieckmann J, Lochynska M, Jaskolski M & Bujacz G (2012) High-resolution structure of *Bombyx mori* lipoprotein 7: crystallographic determination of the identity of the protein and its potential role in detoxification. *Acta Crystallogr* **D68**, 1140–1151.

117 Wojtkowiak A, Witek K, Hennig J & Jaskolski M (2012) Two high-resolution structures of potato endo-1,3-beta-glucanase reveal subdomain flexibility with implications for substrate binding. *Acta Crystallogr* **D68**, 713–723.

118 Weichenberger CX, Pozharski E & Rupp B (2013) Visualizing ligand molecules in twilight electron density. *Acta Crystallogr* **F69**, 195–200.

119 Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wahlby A & Jones TA (2004) The Uppsala Electron-Density Server. *Acta Crystallogr* **D60**, 2240–2249.

120 Cereto-Massague A, Ojeda MJ, Joosten RP, Valls C, Mulero M, Salvado MJ, rola-Arnal A, Arola L, Garcia-Vallve S & Pujadas G (2013) The good, the bad and the dubious: VHELIBS, a validation helper for ligands and binding sites. *J Cheminform* **5**, 36.

121 Hendrickson WA, Strandberg BE, Liljas A, Amzel LM & Lattman EE (1983) True identity of a diffraction pattern attributed to valyl tRNA. *Nature* **303**, 195–196.

122 Janssen BJ, Read RJ, Brunger AT & Gros P (2007) Crystallography: crystallographic evidence for deviating C3b structure. *Nature* **448**, E1–E2.

123 Rupp B (2012) Detection and analysis of unusual features in the structural model and structure-factor data of a birch pollen allergen. *Acta Crystallogr* **F68**, 366–376.

124 Chapman MS, Suh SW, Curmi PM, Cascio D, Smith WW & Eisenberg DS (1988) Tertiary structure of plant RuBisCO: domains and their contacts. *Science* **241**, 71–74.

125 Knight S, Andersson I & Branden CI (1989) Reexamination of the three-dimensional structure of the small subunit of RuBisCo from higher plants. *Science* **244**, 702–705.

126 Dawson RJ & Locher KP (2006) Structure of a bacterial multidrug ABC transporter. *Nature* **443**, 180–185.

127 Navia MA, Fitzgerald PM, McKeever BM, Leu CT, Heimbach JC, Herber WK, Sigal IS, Darke PL & Springer JP (1989) Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. *Nature* **337**, 615–620.

128 Miller M, Jaskólski M, Rao JKM, Leis J & Wlodawer A (1989) Crystal structure of a retroviral protease proves relationship to aspartic protease family. *Nature* **337**, 576–579.

129 Wlodawer A, Miller M, Jaskólski M, Sathyanarayana BK, Baldwin E, Weber IT, Selk LM, Clawson L, Schneider J & Kent SBH (1989) Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. *Science* **245**, 616–621.

130 Hanson MA, Oost TK, Sukonpan C, Rich DH & Stevens RC (2000) Structural basis for BABIM inhibition of botulinum neurotoxin type B protease. *J Am Chem Soc* **122**, 11268–11269.

131 Rupp B & Segelke B (2001) Questions about the structure of the botulinum neurotoxin B light chain in complex with a target peptide. *Nat Struct Biol* **8**, 663–664.

132 Laskowski RA, MacArthur MW, Moss DS & Thornton JM (1993) PROCHECK: program to check the stereochemical quality of protein structures. *J Appl Cryst* **26**, 283–291.

133 Davis IW, Murray LW, Richardson JS & Richardson DC (2004) MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res* **32**, W615–W619.

134 Cooper DR, Porebski PJ, Chruszcz M & Minor W (2011) X-ray crystallography: assessment and validation of protein-small molecule complexes for drug discovery. *Expert Opin Drug Discov* **6**, 771–782.

135 Brzezinski K, Brzuszkiewicz A, Dauter M, Kubicki M, Jaskolski M & Dauter Z (2011) High regularity of Z-DNA revealed by ultra high-resolution crystal structure at 0.55 Å. *Nucleic Acids Res* **39**, 6238–6248.

136 Liebschner D, Dauter M, Brzuszkiewicz A & Dauter Z (2013) On the reproducibility of protein crystal structures: five atomic resolution structures of trypsin. *Acta Crystallogr* **D69**, 1447–1462.