International Tables for Crystallography (2012). Vol. F, Chapter 18.4, pp. 485–498.

Chapter 18.4. Refinement at atomic resolution

Z. Dauter, G. N. Murshudov and K. S. Wilson

18.4.1. The atomic model and a definition of atomic resolution

18.4.1.1. The atomic model

X-rays are diffracted by the electrons that are distributed around the atomic nuclei, and the result of an X-ray crystallographic study is the derived three-dimensional electron-density distribution in the unit cell of the crystal. The elegant simplicity and power of X-ray crystallography arise from the fact that molecular structures are composed of discrete atoms that are treated as spherically symmetric in the usual approximation. This property places such strong restraints on the Fourier transform of the crystal structures of small molecules that the phase problem can be solved by knowledge of the amplitudes alone.

Each atom or ion can be described by up to 11 parameters (Table 18.4.1.1).

The first parameter is the scattering-factor amplitude for the chemical nature of the atom in question, and has been computed and tabulated for all atom types [*International Tables for Crystallography*, Volume C (2004)]. Once the chemical identity of the atom is established, this parameter is fixed.

The next three parameters relate to the positional coordinates of the atom with respect to the origin of the unit cell.

If the resolution is high enough, then the number of observed reflections is sufficient to allow six anisotropic atomic displacement parameters to be used to describe the distribution of the atom positions in different unit cells (Fig. 18.4.1.1). Atomic displacement parameters (ADPs) reflect both the thermal vibration of atoms about the mean position as a function of time (dynamic disorder) and the variation of positions between different unit cells of the crystal arising from its imperfection (static disorder). Contributors to the apparent ADP ($U_{\rm atom}$) can be thought of as follows (Murshudov *et al.*, 1999):

$$U_{\rm atom} = U_{\rm crystal} + U_{\rm TLS} + U_{\rm torsion} + U_{\rm bond}, \qquad (18.4.1.1)$$

where $U_{\rm crystal}$ represents the fact that a crystal itself is in general an anisotropic field that will result in the intensity falling off in an anisotropic manner, $U_{\rm TLS}$ represents a translation/libration/screw (TLS), *i.e.* the overall motion of molecules or domains (Schomaker & Trueblood, 1968; Winn *et al.*, 2001), $U_{\rm torsion}$ is the oscillation along torsion angles and $U_{\rm bond}$ is the oscillation along and across bonds. In principle, all these contributors are highly correlated and it is difficult to separate them from one another. Nevertheless, an understanding of how $U_{\rm atom}$ is a sum of these different components makes it possible to apply atomic anisotropy parameters at different resolutions in a different manner. For example, $U_{\rm crystal} + U_{\rm TLS}$ can be applied at any resolution, as their refinement increases the number of parameters by at most five for $U_{\rm crystal}$ and 20 per independent moiety for $U_{\rm TLS}$. In contrast, refinement of the third contributor does pose a problem, as there is strong correlation between different torsion angles. As an alternative, ADPs along the internal degrees of freedom could in principle be refined. The fourth and final contributor, $U_{\rm bond}$, can only be refined at very high resolution. In real applications, $U_{\rm crystal}$ and $U_{\rm TLS}$ are separated for convenient description of the system, but in practice their effects are indistinguishable.

Table 18.4.1.1
The parameters of an atomic model

Parameter type	Number	Variable or fixed
Atom type	1	Fixed after identification
Positional (x, y, z)	3	Variable, subject to restraints
ADPs:		
isotropic	1	Variable beyond about 2.5 Å
anisotropic	6	Variable beyond about 1.5 Å
Occupancy	1	Variable for visible disorder

In the special case when the tensor $U_{\rm atom}$ is isotropic, *i.e.*, all non-diagonal elements are equal to zero and all diagonal terms are equal to each other, then the atom itself appears to be isotropic and its ADP can be described using only one parameter, $U_{\rm iso}$.

Thus, for a full description of a crystal structure in which all atoms only occupy a single site, nine parameters per atom must be determined: three positional parameters and six anisotropic ADPs. This assumes that the spherical-atom approximation applies and ignores the so-called deformation density resulting from the non-spherical nature of the outer atomic and molecular orbitals involved in the chemical interactions between the atoms (Coppens, 1997).

For disordered regions or features, where atoms can be distributed over two or more identifiable sites, the occupancy introduces a tenth variable for each atom. In many cases, the fractional occupancies are not all independent, but are rather constant for sets of covalently or hydrogen-bonded atoms or for those in non-overlapping solvent networks. This would apply, for example, to partially occupied ligands or side chains with two conformations.

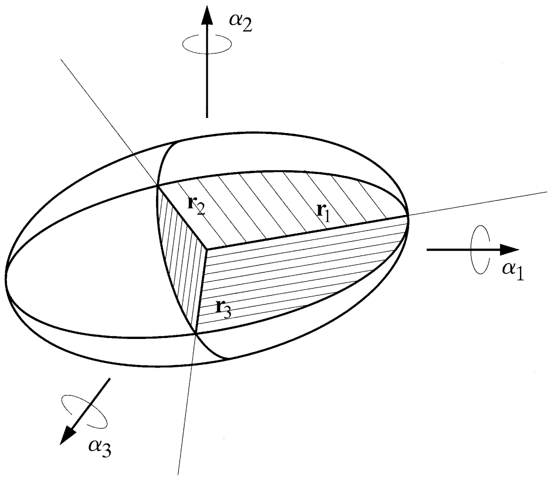

Figure 18.4.1.1
The thermal-ellipsoid model used to represent anisotropic atomic displacement, with major axes indicated. The ellipsoid is drawn with a specified probability of finding an atom inside its contour. Six parameters are necessary to describe the ellipsoid: three represent the dimensions of the major axes and three the orientation of these axes. These six parameters are expressed in terms of a symmetric U tensor and contribute to atomic scattering through the term $\exp[-2\pi^2(U_{11}h^2a^{*2} + U_{22}k^2b^{*2} + U_{33}l^2c^{*2} + 2U_{12}hka^*b^*\cos\gamma^* + 2U_{13}hla^*c^*\cos\beta^* + 2U_{23}klb^*c^*\cos\alpha^*)]$.

Copyright © 2012 International Union of Crystallography

18.4.1.2. What is 'atomic resolution'?

Atomicity is the great simplifying feature of crystallography in terms of structure solution and refinement. For a small-molecule structure, accurate X-ray data usually extend to 0.8 Å, and this has three important implications for crystallography.

(1) *Ab initio phasing using direct methods*. Automatic *ab initio* solution of the phase problem depends on the assumption of positivity and atomicity of the electron density. The fact that current *ab initio* methods in the absence of heavy atoms are only effective when meaningful data extend beyond 1.2 Å reinforces the idea that this is a reasonable working criterion for its definition as atomic resolution. In addition, approaches such as solvent flattening and automated map interpretation benefit enormously from such data.

(2) *Resolved atomic peaks in the Fourier maps*. Although some individual peaks can be seen at resolutions beyond ~2.0 Å, they become more fully resolved at around 1.2 Å.

(3) *Refinement of a full anisotropic model*. The number of reflections is sufficient for the minimization of the discrepancy between the experimentally determined amplitudes or intensities of the Bragg reflections and those calculated from the atomic model with up to ten (usually nine) independent parameters per atom. This has been classically achieved by least-squares refinement as described in *International Tables for Crystallography* Volume C, Chapter 8.1 (Prince & Boggs, 2004) or more recently by maximum-likelihood procedures (Bricogne & Irwin, 1996; Pannu & Read, 1996; Murshudov *et al.*, 1997). For small-molecule structures, accurate amplitude data are normally available to around 0.8 Å, giving an observation-to-parameter ratio of about seven for non-centrosymmetric crystals, which allows positional parameters to be determined with an accuracy approaching 0.001 Å. This reflects the high degree of order of such crystals, in which the molecules in the lattice are in a close-packed array. In addition the X-ray data are of high quality, with a high $I/\sigma(I)$ ratio (and hence low merging R value) even in the outer resolution shells.

It is now necessary to define what constitutes 'atomic resolution'. A pragmatic approach has been that data extending to 1.2 Å or better with at least 50% of the intensities in the outer shell being higher than 2σ is the acceptable limit (Sheldrick, 1990; Sheldrick & Schneider, 1997), which means that the statistical problem of refinement is overdetermined. This appears to remain a good working definition for refinement applications and indeed has been put on a more solid theoretical basis (Morris & Bricogne, 2003; Morris *et al.*, 2004). However, for application of direct phasing methods it is advantageous to record even a small fraction of significant reflections beyond this cutoff. These outer shells should be included in the refinement procedure with correct maximum-likelihood weights, but they will not significantly improve the effective resolution.

This is rarely achieved for crystals of macromolecules: as of October 2009 around 1250 out of 52 000 crystal structures in the Protein Data Bank (PDB) had a resolution higher than 1.2 Å compared to 157 out of 13 000 in March 2000. Firstly, the large unit-cell volume leads to an enormous number of reflections for which the average intensity is weak compared to those for small molecules (see Table 9.1.1.1 in Chapter 9.1). Secondly, the intrinsic disorder of the crystals further reduces the intensities at high Bragg angles and usually gives a resolution cutoff which is much less than atomic. Thirdly, the large solvent content leads to substantial decay of crystal quality under exposure to the X-ray beam at room temperature. While the secondary damage (resulting from the migration of ions and radicals produced by the primary absorption event) is largely avoided by vitrification of such crystals, the effect of primary damage has become significant on high intensity beamlines (see Section 9.1.12). The upper resolution limit of the data affects all stages of a crystallographic analysis, but especially restricts the features of the model that can be independently refined (Table 18.4.1.2). Solutions to the problem of refining macromolecular structures with a paucity of experimental data evolved during the 1970s and 1980s with the use of either constraints or restraints on the stereochemistry, based on that of known small molecules. With constraints, the structure is simplified as a set of rigid chemical units (Diamond, 1971; Herzberg & Sussman, 1983), whereas using restraints, the observation-to-parameter ratio is increased by introduction of prior chemical knowledge of bond lengths and angles (Konnert & Hendrickson, 1980).

Table 18.4.1.2
Features which can be seen in the electron density at different resolutions

Disordered regions will not necessarily be visible even at these limiting values. Some features should be included even at lower resolutions, *e.g.* hydrogen atoms at their riding positions can be incorporated at 2.0 Å, but their positions will not be verifiable from the density. The contents of this table should not be taken as dogmatic rules, but as approximate guidelines.

Resolution (Å)	Feature
0.8	Deformation density, *i.e.* deviation from the spherical-atom model
1.0	Hydrogen atoms
1.5	Anisotropic atomic displacement
2.0	Multiple conformations
2.5	Individual isotropic atomic displacement
3.5	Overall temperature factor
4.0	α-Helices and β-sheets
6.0	Domain envelopes

As expected, atoms with different ADPs contribute differently to the diffraction intensities, as discussed by Cruickshank (1999*a*,*b*). The relative contribution of the different atoms to a given reflection depends on the difference between their ADPs $\{\exp[-(B_1 - B_2)s^2]$, where $s = \sin\theta/\lambda\}$. Clearly, if the average ADP of a molecule is small, then the spread will also be narrow, and most atoms will contribute to diffraction over the whole range of resolution. When the mean ADP is large, then the spread of the ADPs will be wide, and fewer atoms will contribute to the high-resolution intensities (Fig. 18.4.1.2).

Three advances in experimental techniques have combined effectively to overcome these problems for an increasing number of well ordered macromolecular crystals, namely the use of high-intensity synchrotron radiation (SR), efficient two-dimensional detectors and cryogenic freezing (discussed in Parts 8, 7 and 10, respectively). These advances mean that there is no longer a sharp division between small-molecule and macromolecular crystallography, but rather a continuum from small through medium-sized structures, such as cyclodextrins and other supramolecules, to proteins. The inherent disorder in the crystal generally increases with the size of the structure, due in part to the increasing solvent content. Thus, it has become tractable to refine a significant number of protein structures at atomic resolution with a full anisotropic model (Dauter, Lamzin & Wilson, 1997; Dauter, 2003). This work of course benefits tremendously from the experience and algorithms of small-molecule crystallography, but does pose special problems of its own. The tech-

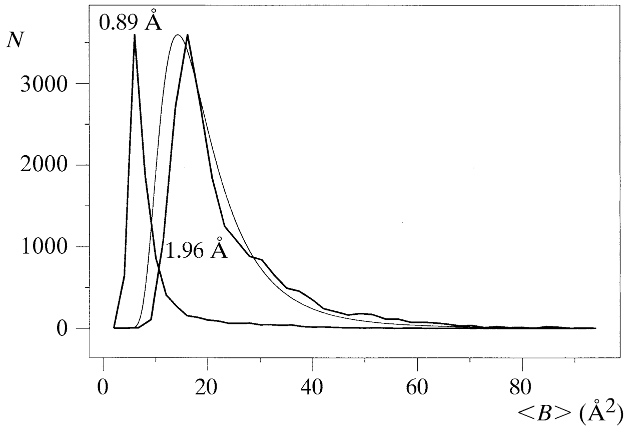

Figure 18.4.1.2
Histograms of B values for a protein structure, *Micrococcus lysodecticus* catalase (Murshudov *et al.*, 1999), for two different crystals which diffracted to different limiting resolutions. For both crystals, the resolution cutoff reflects the real diffraction limit from the sample, and hence its level of order. At 0.89 Å, the mean B value is 8.3 Å^2 and the width of the distribution is small. In contrast, at 1.96 Å, the mean B value is 25.5 Å^2 and the spread is correspondingly large. Thus, for the 0.89 Å crystal, most atoms contribute to the high-resolution terms, whereas for the 1.96 Å crystal, only the atoms with lower B values do so. The thin line shows the theoretical inverse gamma distribution $\mathrm{IG}(B) = (b/2)^{d/2}/\Gamma(d/2)B^{-(d+2)/2}\exp[-b(2B)]$, where b and d are the parameters of the distribution, and Γ is the gamma function. For this figure, the values $b = 2$ and $d = 10$ were chosen, which correspond to a mean B value of 20 Å^2 and σ_B of 11 Å^2. In the gamma distribution, the abscissa was multiplied by $8\pi^2$ to make it comparable with the measured B values. All three histograms were normalized to the same scale.

niques of solving and refining macromolecular structures thus also overlap with those conventionally used for small molecules; a prime example is the use of *SHELXL* (Sheldrick, 2008), which was developed for small structures and has now been extended to treat macromolecules.

18.4.1.3. A theoretical approach to 'atomic resolution'

An alternative and stricter definition of atomic resolution comes from using a measure of the information content of the data. There are a variety of definitions of the information in the data about the postulated model (see, for example, O'Hagan, 1994). A suitable one is the Bayesian definition for quadratic information measure:

$$I_Q(p, F) = \mathrm{tr}(A\{\mathrm{var}(p) - E[\mathrm{var}(p, F)]\}), \qquad (18.4.1.2)$$

where I_Q is the quadratic information measure, p is the vector of parameters, F is the experimental data, var(p) is the variance matrix corresponding to prior knowledge, var(p, F) is the variance matrix corresponding to the posterior distribution (which includes prior knowledge and likelihood), E is the expectation, tr is the trace operator (*i.e.* the sum of the diagonal terms of the matrix) and A is the matrix through which the relative importance of different parameters or combinations of parameters is introduced. For example, if A is the identity matrix, then the information measure is unitary and all parameters are assigned the same weight. If A is the identity matrix for positional parameters and zero for ADPs, then only the information about positional parameters is included. By appropriate choice of A, the information about selected key features, such as the active site, can be estimated.

Equation (18.4.1.2) shows how much the experiment reduces the uncertainty in given parameters. Prior knowledge is usually taken to be information about bond lengths, bond angles and other chemical features of the molecule, known before the experiment has been carried out. In the case of an experiment designed to provide information about the ligated protein or mutant, when information about differences between two (or more) different states is needed, the prior knowledge can be thought of in a different way – as knowledge about the native protein.

Unfortunately, there are problems in applying equation (18.4.1.2). Firstly, careful analysis of the prior knowledge and its variance is essential. The target values used at present, or more properly the distributions for these values, need to be re-evaluated. Another problem concerns the integration required to compute the expectation value (E). Nevertheless, the equation provides some idea of how much information about a postulated model can be extracted from a given experiment.

This alternative definition of atomic resolution assumes that the second term of equation (18.4.1.2) for positional parameters is sufficiently close to zero for most atoms to be resolved from all their neighbours. Defining atomic resolution using this information measure reflects the importance of both the quality and quantity of the data [through the posterior var(p, F)]. In addition, data may come from more than one crystal, in which case the information will be correspondingly increased. There may be additional data from mutant and/or complexed protein crystals, where, again, the information measure will be increased and, moreover, the differences between different states can be analysed. The effect of redundancy of different crystals of the same molecule(s) in different space groups is to reduce the limit of data necessary for achieving atomic resolution, which is equivalent to the advantage of noncrystallographic averaging.

Thus, in practice, while it would be ideal to develop the strict application of equation (18.4.1.2), for the present it is necessary to rely on the pragmatic approach in Section 18.4.1.2.

18.4.2. Data

The quality of the refined model relies finally on that of the available experimental data. Data collection has been covered extensively in Chapter 9.1 and will not be discussed here.

18.4.2.1. Data quality

As can be seen from equation (18.4.1.2), the measure of information about all or part of the crystal contents depends strongly on the quality and quantity of the data. Of course, before the experiment is carried out some questions should be answered. Firstly, what is the aim of the experiment? Secondly, what is the cost of the experiment and what are the available resources? With modern techniques, if SR is used with an efficient detector, the cost of the experiment for different resolutions does not vary greatly (provided that a suitable quality crystal is available). In practice, the apparent increase in cost to attain high-resolution data will generally make solving the phase problem both easier and faster. A full analysis at atomic resolution provides a wealth of additional structural detail which may shed light on the subtleties of the protein's chemistry not seen at lower resolution. However, this may require some considerable time and effort, and is an area where development of more automated approaches would be beneficial. In contrast, low-resolution data can make it difficult to answer not only the question currently being asked, but can also necessitate further experiments to address other problems that arise.

While the information content of the data appears to depend quantitatively on the nominal resolution, in fact it is dependent on the data quality throughout the resolution range, and both high- and low-resolution completeness and their statistical significance affect the information content of the data and derived model. High-intensity low-resolution terms remain important for refinement at atomic resolution, as they define the contrast in the density maps between solvent and protein, and because their omission biases the refinement, especially that of parameters such as the ADPs. To judge the effective resolution of the diffraction data set, the concept of 'optical resolution', which can be estimated from the shape of the origin peak in the Patterson synthesis, may be very useful (Vaguine *et al.*, 1999).

The rejection of low-intensity observations will also introduce bias. In particular, all the maps calculated for visual or computer inspection by Fourier transformation are diminished in quality by omission of any terms, but are especially affected by omission of strong low-resolution data. This is particularly true in the early stages of structure solution, where low-resolution data can be vital. Although most phase-improvement algorithms rely on relations between all reflections, terms involving low-resolution reflections will be large, will be involved in many relations and will play a dominant role. Hence, omission of these terms will severely degrade the power of these methods, which may indeed converge to solutions that have nothing whatsoever to do with the real structure.

18.4.2.2. Anisotropic scaling

The intensity data from a crystal may display anisotropy, *i.e.*, the intensity fall-off with resolution will vary with direction, and may be much higher along one crystal axis than along another. If the structure is to be refined with an isotropic atomic model (either because there are insufficient data or the programs used cannot handle anisotropic parameters), then the fall-off of the calculated F^2 values will, of necessity, also be isotropic. In this situation, an improved agreement between observed and calculated F^2 values can be obtained either by using anisotropic scaling during data reduction to the expected Wilson distribution of intensities, or by including a maximum of six overall anisotropic parameters during refinement. This will result in an isotropic set of F^2 values. For crystals with a high degree of anisotropy in the experimental data, this can lead to a substantial drop of several per cent in R and $R_{\rm free}$ (Sheriff & Hendrickson, 1987; Murshudov *et al.*, 1998).

This ambiguity effectively disappears with use of an anisotropic atomic model. The individual ADPs, including contributions from both static and thermal disorder, take up relative individual displacements, but also the overall anisotropy of the experimental F^2 values. The significance of the overall anisotropy is a point of some contention, and its physical meaning is not clear. It may represent asymmetric crystal imperfection or anisotropic overall displacement of molecules in the lattice related to TLS parameters. Refinement of TLS parameters, which can be performed using, for example, *RESTRAIN* (Driessen *et al.*, 1989) *REFMAC* (Winn *et al.*, 2001) or *PHENIX.REFINE* (Terwilliger *et al.*, 2008), removes the overall crystal contribution to the ADPs.

It is important that at least the intensity or amplitude data be deposited in the PDB as measured, without any anisotropic correction being applied. For the refined model, complete information on any overall anisotropic and TLS modelling must be explicitly included, as well as the individual atomic ADPs.

18.4.3. Computational algorithms and strategies

18.4.3.1. Classical least-squares refinement of small molecules

The principles of the least-squares method of minimization are described in *International Tables for Crystallography* Volume C (2004). Least squares involves the construction of an order $N \times N$ normal matrix, where N is the number of parameters, representing a system of least-squares equations, whose solution provides estimates of adjustments to the current atomic parameters. The problem is nonlinear and the matrix construction and solution must be iterated until convergence is achieved. In addition, inversion of the matrix at convergence provides an approximation to the standard uncertainties for each individual parameter refined according to the Cramer–Rao inequality (Stuart *et al.*, 1999). Indeed, this is the only method available so far that gives such estimates properly.

However, even for small molecules there may be some disordered regions which will require the imposition of restraints, as is the case for macromolecules (see below), and the presence of such restraints means that the error estimates no longer reflect the information from the X-ray data alone. If the problem of how restraints affect the error estimates could be resolved, then inversion of the matrix corresponding to the second derivative of the posterior distribution would provide standard uncertainties incorporating both the prior knowledge, such as the restraints, and the experimental data. Equation (18.4.1.2) for information measure could then be applied, but this requires further development. For small structures, the speed and memory of modern computers have reduced the requirements for such calculations to the level of seconds, and the computational requirements form a trivial part of the structure analysis. Recent developments in the application of fast Fourier transform methods to normal matrix–vector multiplication (Strokopytov, 2008) and fast information matrix evaluation (Steiner *et al.*, 2003) suggest that fast calculations for macromolecular structures may be available in the foreseeable future.

18.4.3.2. Least-squares refinement of large structures

The size of the computational problem increases dramatically with the size of the unit cell, as the number of terms in the matrix increases with the square of the number of parameters. Furthermore, construction of each element depends on the number of reflections. For macromolecular structures, computation of a full matrix is at present prohibitively expensive in terms of CPU time and memory. A variety of simplifying approaches have been developed, but all suffer from a poorer estimate of the standard uncertainties and from a more limited range and speed of convergence.

The first is the block-matrix approach, where instead of the full matrix, only square blocks along the matrix diagonal are constructed, involving groups of parameters that are expected to be correlated. The correlation between parameters belonging to different blocks is therefore neglected completely. In this way, the whole least-squares minimization is split into a set of smaller independent units. In principle this leads to the same solution, but more slowly and with less precise error estimates. Nevertheless, block-matrix approaches remain essential for tractable matrix inversion for macromolecular structures.

A further simplification involves the conjugate-gradient method or the diagonal approximation to the normal matrix (the second derivative of minus the log of the likelihood function in the case of maximum likelihood), which essentially ignores all off-diagonal terms of the least-squares matrix. For the conjugate-gradient approach, all diagonal terms of the matrix are equal. However, the range and speed of convergence are substantially reduced, and standard uncertainties can no longer be estimated directly by matrix inversion.

18.4.3.3. Fast Fourier transform

Conventional least-squares programs use the structure-factor equation and associated derivatives, with the summation extending over all atoms and all reflections. This is immensely slow in computational terms for large structures, but it has the advantage of providing precise values.

An alternative procedure, where the computer time is reduced from being proportional to N^2 to $N \log N$, involves the use of fast Fourier algorithms for the computation of structure factors and derivatives (Ten Eyck, 1973, 1977; Agarwal, 1978). This can involve some interpolation and the limitation of the volume of electron-density maps to which individual atoms contribute. Such algorithms have been exploited extensively in macromolecular refinement programs such as *PROLSQ* (Konnert & Hendrickson, 1980), *XPLOR* (Brünger, 1992*b*), *TNT* (Tronrud, 1997), *RESTRAIN* (Driessen *et al.*, 1989), *REFMAC* (Murshudov *et al.*, 1997), *CNS* (Brünger *et al.*, 1998) and *PHENIX.REFINE* (Terwilliger *et al.*, 2008), but have largely been restricted to the diagonal approximation. *XPLOR* and *CNS* use the conjugate-gradient method, which relies only on the first derivatives and ignores the second derivatives. In all other programs, the diagonal approximation is used for the second-derivative matrix.

18.4.3.4. Maximum likelihood

This provides a statistically sounder alternative to least squares, especially in the early stages of refinement when the model lies far from the minimum. This approach increases the radius of convergence, takes into account experimental uncertainties, and in the final stages gives results similar to least squares but with improved weights (Bricogne & Irwin, 1996; Murshudov *et al.*, 1997). The maximum-likelihood approach has been extended to allow refinement of a full atomic anisotropic model while retaining the use of fast Fourier algorithms (Murshudov *et al.*, 1999). A remaining limitation is the use of the diagonal approximation, which prevents the computation of standard uncertainties of individual parameters. Algorithms that will alleviate this limitation can be foreseen, and they are expected to be implemented in the future.

18.4.3.5. Twinning

A non-negligible fraction of protein crystals turn out to be merohedrally twinned (Lebedev *et al.*, 2006), which requires special treatment of the diffraction data and proper treatment of this phenomenon during structure solution and more especially refinement (Yeates, 1997; Chapters 18.11 and 18.12). The twinning problem can be approached in two distinctly different ways. In the first, the data are explicitly 'detwinned' to produce an adjusted set of amplitudes [for example using the UCLA detwinning server (http://nihserver.mbi.ucla.edu/Twinning/) or the *DETWIN* program (Collaborative Computational Project, Number 4, 1994)], which are then subjected to conventional refinement. However, detwinning can only be successfully applied if the twinning fraction is not too high, since the error in the resulting amplitudes increases as the fraction approaches 50%. A second, and certainly preferred, approach is to include the twinning fraction as a variable during refinement. This has been implemented in *SHELXL*, *PHENIX.REFINE* and *REFMAC*, and is widely used. In *CNS*, it is possible to set the twin fraction, but not to refine it. This is of special relevance for atomic resolution structures, as even a small degree of twinning will have a significant effect on the interpretability of fine features.

18.4.3.6. Computer power

There are no longer any restrictions on the full-matrix refinement of small-molecule crystal structures. However, the large size of the matrix, which increases as N^2, where N is the number of parameters, means that for macromolecules with thousands of independent atoms this approach is intractable with the computing resources normally available to the crystallographer. By extrapolating the progress in computing power experienced in recent years, it can be envisaged that the limitations will disappear during the next decade, as those for small structures have disappeared since the 1960s. Indeed, the advances in the speed of CPUs, computer memory and disk capacity continue to transform the field.

18.4.4. Computational options and tactics

18.4.4.1. Use of F (amplitudes) or F^2 (intensities)

The X-ray experiment provides two-dimensional diffraction images. These are transformed to integrated but unscaled data, which are transformed to Bragg reflection intensities that are subsequently transformed to structure-factor amplitudes. At each transformation some assumptions are used, and the results will depend on their validity. Invalid assumptions will introduce bias toward these assumptions into the resulting data. Ideally, refinement (or estimation of parameters) should be against data that are as close as possible to the experimental observations, eliminating at least some of the invalid assumptions. Extrapolating this to the extreme, refinement should use the images as observable data, but this poses several severe problems, depending on data quantity and the lack of an appropriate statistical model.

Alternatively, the transformation of data could be improved by revising the assumptions. The intensities are closer to the real experiment than are the structure-factor amplitudes, and use of intensities would reduce the bias. However, there are some difficulties in the implementation of intensity-based likelihood refinement (Pannu & Read, 1996).

Gaussian approximation to intensity-based likelihood (Murshudov *et al.*, 1997) would avoid these difficulties, since a Gaussian distribution of error can be assumed in the intensities but not the amplitudes. However, errors in intensities may not only be the result of counting statistics, but may have additional contributions from factors such as crystal disorder and motion of the molecules in the lattice during data collection.

Nevertheless, the problem of how to treat weak reflections remains. Some of the measured intensities will be negative, as a result of statistical errors of observation, and the proportion of such measurements will be relatively large for weakly diffracting macromolecular structures, especially at atomic resolution. This is less important for intensity-based likelihood than for the amplitude-based approach. French & Wilson (1978) have given a Bayesian approach for the derivation of structure-factor amplitudes from intensities using Wilson's distribution (Wilson, 1942) as a prior, but there is room for improvement in this approach. Firstly, the Wilson distribution could be upgraded using the scaling techniques suggested by Blessing (1997) and Cowtan &

Main (1998), and secondly, information about effects such as pseudosymmetry could be exploited.

Another argument for the use of intensities rather than amplitudes is relevant to least squares, where the derivative for amplitude-based refinement with respect to $F_{\rm calc}$ is singular when $F_{\rm calc}$ is equal to zero (Schwarzenbach *et al.*, 1995). This is not the case for intensity-based least squares. In applying maximum likelihood, this problem does not arise (Pannu & Read, 1996; Murshudov *et al.*, 1997).

Finally, while there may be some advantages in refining against intensities, Fourier syntheses always require structure-factor amplitudes.

18.4.4.2. Restraints on coordinates and ADPs

For a good small-molecule crystal the experimental X-ray data extend to ~0.8 Å spacing and the structure can be refined against the X-ray data alone. The resulting accuracy of the atomic coordinates will generally be better than 0.01 Å. However, even for small-molecule structures, disordered regions require the imposition of stereochemical restraints (or constraints) if the chemical integrity is to be preserved and the ADPs are to be realistic.

The typical situation for protein crystals is quite different, with atomic resolution being the exception rather than the rule. Thus, for proteins the geometry of the atomic model needs to be restrained, both in terms of geometry and ADPs. The geometric target values have been established from a set of amino-acid and small-peptide structures (see Chapter 18.3 by Engh & Huber), for which the bond-length r.m.s.d. is about 0.02 Å. In the present context, we restrict the discussion to bond lengths, but this is representative of the other restraints. Clearly, the relative contribution of the X-ray data and the restraints on the final parameters varies as a function of resolution. The restraints dominate at low resolution, while by the time 0.8 Å spacing is achieved, the restraints will be essentially irrelevant for well ordered regions. The imposition of bond-length restraints with target deviations of ~0.02 Å means that the distribution of bond lengths in the final model will be of the same order, independent of the resolution of the X-ray data. However, this must not be taken to imply that the accuracy of the atomic parameters is invariant with the overall resolution and, more importantly, the atomic displacement. To be explicit, the accuracy of the atomic positions decreases (1) as the resolution becomes worse (*i.e.*, the number of X-ray observations decreases) and (2) as the ADPs become larger. While this should be obvious to the practicing crystallographer, it may not be so apparent to the less expert user of the PDB (Wlodawer *et al.*, 2008).

Jaskolski *et al.* (2007) analysed ten structures from the PDB refined at ultra-high (better than 0.8 Å) resolution to investigate appropriate geometrical restraints. They confirmed the general correctness of values in the Engh & Huber dictionary and showed that the mean observed deviations in these ten structures from the target bond lengths were indeed roughly 0.02 Å. They therefore postulated that 0.02 Å was an appropriate value to use in applying stereochemical restraints to protein structures in general. There has been some dispute about this value (Tickle, 2007) but we believe it to be appropriate. A more detailed analysis of this issue has been performed, suggesting that target values differ depending on the structural context (Karplus *et al.*, 2008); this may lead to some fine adjustments in the target dictionary, as predicted earlier (EU 3-D Validation Network, 1998).

In analogy to the geometrical restraints based on the Engh & Huber dictionary, anisotropic ADP restraints were established in the *SHELX* program suite (Sheldrick, 2008). For example, they prevent atoms from becoming unrealistically anisotropic and restrain the shapes of the ellipsoids of bonded atoms to be not too dissimilar. Riding hydrogen atoms are assigned constrained isotropic ADPs based on those of the parent atoms.

A more theoretical justification for use of restraints is that refinement can be considered as Bayesian estimation. From this point of view, all available and usable prior knowledge should be exploited, as it should not harm the parameter estimation during refinement. Bayesian estimation shows asymptotic behaviour (Box & Tiao, 1973), *i.e.*, when the number of observations becomes large, the experimental data override the prior knowledge. In this sense, the purpose of the experiment is to enhance our knowledge about the molecule, and the procedure should be cumulative, *i.e.*, the result of the old experiment should serve as prior knowledge for the design and treatment of new experiments (Box & Tiao, 1973; Stuart *et al.*, 1999; O'Hagan, 1994). However, there are problems in using restraints. For example, the probability distribution reflecting the degree of belief in the restraints is not good enough. Use of a Gaussian approximation to distributions of distances, angles and other geometric properties has not been justified. Firstly, the distribution of geometric parameters depends strongly on ADPs. Secondly, different geometric parameters are correlated. Thirdly, many geometric parameters (*e.g.* bond angles, torsion angles) are dependent on the conformation, configuration and environment of the molecule in question (Karplus *et al.*, 2008; Gelbin *et al.*, 1996). This problem should be the subject of further investigation.

In summary, the atomic resolution structures to date confirm that a mean deviation of bond lengths from target values of 0.02 Å (and comparable values for other restraint types) is appropriate, but may be subject to minor adjustments. These levels of restraint should be applied at all resolutions: the stereochemistry should be neither over- nor under-restrained.

18.4.4.3. Partial occupancy

It may be necessary to refine one additional parameter, the occupancy factor of an atomic site, for structures possessing regions that are spatially or temporally disordered, with some atoms lying in more than one discrete site. The sum of the occupancies for alternative individual sites of a protein atom must be 1.0.

For macromolecules, the occupancy factor is important in several situations, including the following:

(1) when a protein or ligand atom is present in all molecules in the lattice, but can lie in more than one position due to alternative conformations;

(2) for the solvent region, where there may be overlapping and mutually exclusive solvent networks;

(3) when ligand-binding sites are only partially occupied due to weak binding constants, and the structures represent a mixture of native enzyme with associated solvent and the complex structure;

(4) when there is a mixture of protein residues in the crystal, due to inhomogeneity of the sample arising from polymorphism, a mixture of mutant and wild-type protein, or other causes.

Unfortunately, the occupancy parameter is highly correlated with the ADP, and it is difficult to model these two parameters at resolutions less than atomic. Even at atomic resolution, it can

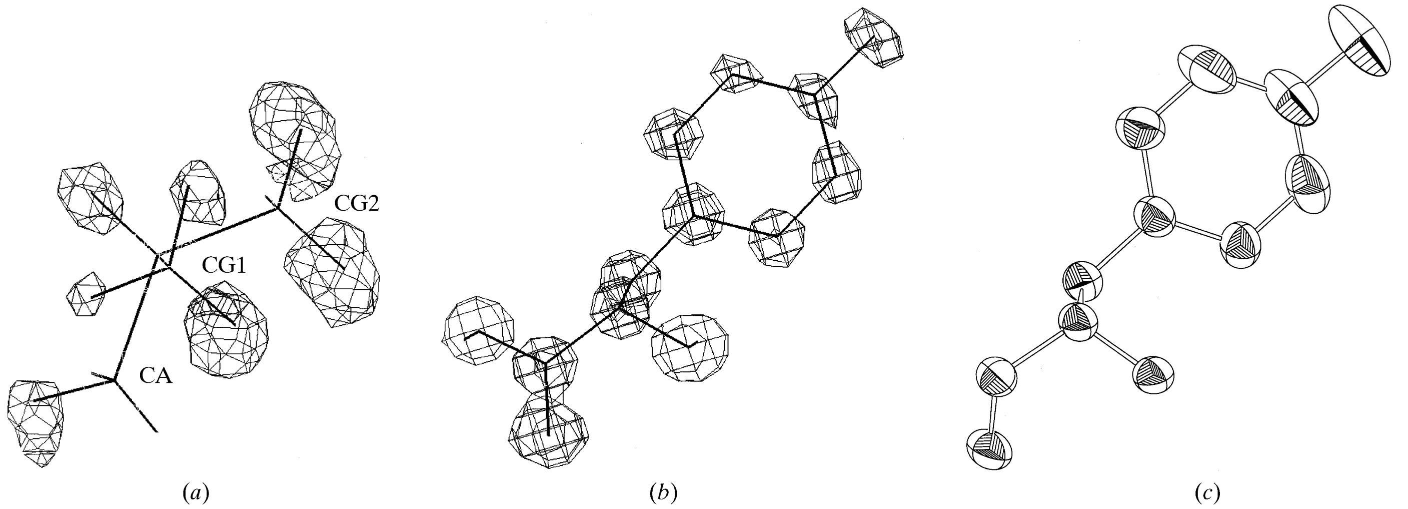

Figure 18.4.5.1
(*a*), (*b*) Representative electron-density maps for the refinement of *Clostridium acidurici* ferredoxin at 0.94 Å resolution (Dauter, Wilson *et al.*, 1997). (*a*) The density for hydrogen atoms (at 3σ) omitted from the structure-factor calculation for Val42. (*b*) The $(2F_o - F_c)$ density for Tyr30, contoured at 3σ. (*c*) The thermal ellipsoids corresponding to (*b*), drawn at the 33% probability level using *ORTEP*II (Johnson, 1976). There is a clear correlation between the density in (*b*) and the ellipsoids in (*c*), showing increased displacement towards the end of the side chain, particularly in the plane of the phenyl ring.

prove difficult to refine the occupancy satisfactorily with statistical certainty.

18.4.4.4. Validation of extra parameters during the refinement process

The introduction of additional parameters into the model always results in a reduction in the least-squares or maximum-likelihood residual – in crystallographic terms, the *R* factor. However, the statistical significance of this reduction is not always clear, since this simultaneously reduces the observation-to-parameter ratio. It is therefore important to validate the significance of the introduction of further parameters into the model on a statistical basis.

Brünger (1992*a*) introduced the concept of statistical cross validation to evaluate the significance of introducing extra features into the atomic model. For this, a small and randomly distributed subset of the experimental observations is excluded from the refinement procedure, and the residual against this subset of reflections is termed $R_{\rm free}$. It is generally sufficient to include about 1000 reflections in the $R_{\rm free}$ subset; further increase in this number provides little, if any, statistical advantage but diminishes the power of the minimization procedure. For atomic resolution structures, cross validation is important in establishing whether the introduction of an additional type of feature to the model (with its associated increase in parameters) is justified. There are two limitations to this. Firstly, if $R_{\rm free}$ shows zero or minimal decrease compared to that in the *R* factor, the significance remains unclear. Secondly, the introduction of individual features, for example the partial occupancy of five water molecules, can provide only a very small change in $R_{\rm free}$, which will be impossible to substantiate. To recapitulate, at atomic resolution the prime use of cross validation is in establishing protocols with regard to extended sets of parameter types. The sets thus defined will depend on the quality of the data.

In the final analysis, validation of individual features depends on the electron density, and Fourier maps must be judiciously inspected. Nevertheless, this remains a somewhat subjective approach and is in practice intractable for extensive sets of parameters, such as the occupancies and ADPs of all solvent sites. For the latter, automated procedures, which are being developed at present, are an absolute necessity, but they may not be optimal in the final stages of structure analysis, and visual inspection of the model and density is often needed.

The problems of limited data and reparameterization of the model remain. At high resolution, reparameterization means having the same number of atoms, but changing the number of parameters to increase their statistical significance, for example switching from an anisotropic to an isotropic atomic model or *vice versa*. In contrast, when reparameterization is applied at low resolution, this usually involves constraints, *i.e.*, a reduction in the number of independent atoms, but this is not an ideal procedure, as real chemical entities of the model are sacrificed. Reducing the number of independent atoms will inevitably result in disagreement between the experiment and model, which in turn will affect the precision of other parameters. It would be more appropriate to reduce the number of parameters without sacrificing the number of atoms, for example by describing the model in torsion-angle space. Water poses a particular problem, as at low as well as at high resolution the water molecules cannot all be described as discrete atoms. Algorithms are needed to describe them as a continuous model with only a few parameters. In the simplest model, the solvent can be described as a constant electron density.

18.4.5. Features in the refined model

All features of the refined model are more accurately defined if the data extend to higher resolution (Fig. 18.4.5.1). In this section, those features that are especially enhanced in an atomic resolution analysis are described. Introduction of an additional feature to the model should be assessed by the use of cross- or self-validation tools: only then can the significance of the parameters added to the model be substantiated.

18.4.5.1. Hydrogen atoms

Hydrogen atoms possess only a single electron and therefore have low electron density and are relatively poorly defined in X-ray studies. They play central roles in the function of proteins, but at the traditional resolution limits of macromolecular structure analyses their positions can only be inferred rather than defined from the experimental data. Indeed, even at a resolution of 2.5 Å, hydrogen atoms should be included in the refined model, as their exclusion biases the position of the heavier atoms, but with their 'riding' positions fixed by those of the parent atoms.

As for small structures, independent refinement of hydrogen-atom positions is not always warranted, even by atomic resolution data, and hydrogen atoms are rather attached as riding rigidly on the positions of the parent atoms. Nevertheless, atomic resolution data allow the experimental confirmation of the positions of many of the hydrogen atoms in the electron-density maps, as they do account for one-sixth of the diffracting power of a carbon atom. Inspection of the maps can in principle allow the identification of (1) the presence or absence of hydrogen atoms on key residues, such as histidine, aspartate and glutamate or on ligands, and (2) the correct location of hydrogen atoms where more than one position is possible, such as in the hydroxyl groups of serine, threonine or tyrosine.

The correct placement of hydrogen atoms riding on their parent atoms involves computation of the appropriate position after each cycle of refinement. This is done automatically by programs such as *SHELXL*, *REFMAC* or *PHENIX.REFINE*. For rigid groups such as the NH amide, aromatic rings, $-CH_2-$ or $=CH-$, the position is accurately defined by the bonding scheme. For groups such as methyl CH_3 or OH, the position is not absolutely defined, and the software is required to make judgmental decisions. For example, *SHELXL* offers the opportunity to inspect the maximum density on a circular Fourier synthesis for optimal positioning. The bond length is fixed according to results from a small-molecule database. The location of hydrogen atoms on polar atoms can be assisted by software that analyses the local hydrogen-bonding networks; this involves maximization of the hydrogen-bonding potential of the relevant groups. Sheldrick advocates assigning ADP values to riding hydrogen atoms of 1.2 times that of the parent atom, or 1.5 times in the case of methyl and similar entities.

18.4.5.2. Anisotropic atomic displacement parameters

Refinement of an isotropic model involves four independent parameters per atom, three positional and one isotropic ADP. In contrast, an anisotropic model requires nine parameters per atom, with the anisotropic atomic displacement described by an ellipsoid represented by six parameters. At 1 Å resolution, the data certainly justify an anisotropic atomic model. Extension of the model from isotropic to anisotropic should generally result in a reduction in the R factor of the order of 5–6% and a comparable drop in $R_{\rm free}$. As a consequence of the diminution of the observable-to-parameter ratio, the R factor at all resolutions will drop by a similar amount; however, $R_{\rm free}$ will not. Experience shows that at 2 Å or less there is no drop in $R_{\rm free}$, and an anisotropic model is totally unsupported by the data. At intermediate resolutions, the result depends on the data quality and completeness. At lower resolution, to account for anisotropy of the atoms, the overall motion of molecules or domains can be refined using translation/libration/screw (TLS) parameters (Schomaker & Trueblood, 1968).

Until the end of the 1990s, anisotropic ADPs had only been handled by programs originally developed for small-molecule analysis, which use conventional algebraic computations of the calculated structure-factor amplitudes, *SHELXL* being a prime example. A limitation of this approach is the substantial computation time required. The use of fast-Fourier-transform algorithms for the structure-factor calculation leads to a significant saving in time (Murshudov *et al.*, 1999). Anisotropic modelling of the individual ADPs is essential if the thermal vibration is to be analysed in terms of coordinated motion of the whole molecule or of domains (Schomaker & Trueblood, 1968). Painter & Merritt (2006) have provided a means of analysing refined models to suggest appropriate TLS groupings.

18.4.5.3. Alternative conformations

Proteins are not rigid units with a single allowed conformation. *In vivo* they spontaneously fold from a linear sequence of amino acids to provide a three-dimensional phenotype that may exhibit substantial flexibility, which can play a central role in biological function, for example in the induced fit of an enzyme by a substrate or in allosteric conformational changes. Flexibility is reflected in the nature of the protein crystals, in particular the presence of regions of disordered solvent between neighbouring macromolecules in the lattice (see Section 18.4.5.6).

The structure tends to be highly ordered at the core of the protein, or more properly, at the core of the individual domains. Atoms in these regions in the most ordered protein crystals have ADP values comparable to those of small molecules, reflecting the fact that they are, in essence, closely packed by surrounding protein. In general, as one moves towards the surface of the protein, the situation becomes increasingly fluid. Side chains and even limited stretches of the main chain may show two (or multiple) conformations. These may be significant for the biological function of the protein.

The ability to model the alternative conformations is highly resolution dependent. At atomic resolution, the occupancy of two alternative but well defined conformations can be refined to an accuracy of about 5%, thus second conformations can be seen, provided that their occupancy is about 10% or higher. The limited number of proteins for which atomic resolution structures are available suggest that up to 20% of the 'ordered residues' show multiple conformations. This confers even further complexity on the description of the protein model. A constraint can be imposed on residues with multiple conformations: namely that the sum of all the alternatives must be 1.0. Protein regions (whether they are side- or main-chain regions) with alternative conformations and partial occupancy can form clusters in the unit cell with complementary occupancy. This often coincides with alternative sets of solvent sites, which should also be refined with complementary occupancies.

The atoms in two alternative conformations occupy independent and discrete sites in the lattice, about which each vibrates. However, if the spacing between two sites is small and the vibration of each is large, then it becomes impossible to differentiate a single site with high anisotropy from two separate sites. There is no absolute rule for such cases: programs such as *SHELXL* place an upper limit on the anisotropy and then suggest splitting the atom over two sites. Some regions can show even higher levels of disorder, with no electron density being visible for their constituent atoms. Such fully disordered regions do not contribute to the diffraction at high resolution, and the definition of their location will not be improved with atomic resolution data.

18.4.5.4. Ordered solvent water

A protein crystal typically contains ~50% aqueous solvent. This is roughly divided into two separate zones. The first is a set of highly ordered sites close to the surface of the protein. The second, lying remote from the protein surface, is essentially composed of fluid water, with no order between different unit cells.

At room temperature, the solvent sites around the surface are assumed to be in dynamic equilibrium with the surrounding fluid, as for a protein in solution. Nevertheless, the observation of apparently ordered solvent sites on the surface indicates that these are occupied most of the time. The waters are organized in hydrogen-bonded networks, both to the protein and with one another. The most highly ordered water sites lie in the first solvent shell, where at least one contact is made directly to the protein. For the second and subsequent shells the degree of order diminishes: such shells form an intermediate grey level between the ordered protein and the totally disordered fluid. Indeed, the flexible residues on the surface form part of the continuum between a solid and liquid phase.

In the ordered region, the solvent structure can be modelled by discrete sites whose positional parameters and ADPs can be refined. For sites with low ADPs, the refinement is stable and their behaviour is well defined. As the ADPs increase, or more likely the associated occupancy in a particular site falls, the behaviour deteriorates, until finally the existence of the site becomes dubious. There is no hard cutoff for the reality of a weak solvent site. However, the number and significance of solvent sites are increased by atomic resolution data. Despite the fact that the waters contribute only weakly to the high-resolution terms, the improved accuracy of the rest of the structure and the reduction of noise due to the high resolution mean that their positions become better defined.

Indeed, the occupancy of some solvent sites can be refined if the resolution is sufficient, or at least their fractional occupancy can be estimated and kept fixed (Walsh *et al.*, 1998). This leads to the possibility of defining overlapping water networks with alternative hydrogen-bonding schemes. This can be a most time-consuming step in atomic resolution refinement, and a trade-off finally has to be made between the relevance of any improvement in the model and the time spent.

18.4.5.5. Automatic location of water sites

The protein itself has a clearly defined chemical structure, and the number of atoms to be positioned and how they are bonded to one another are known at the start of model building. The solvent region is in marked contrast to this, as the number of ordered water sites is not known *a priori*, and the distances between them are less well defined, their occupancy is uncertain, and there may be overlapping networks of partially occupied solvent sites. Those of low occupancy lie at the level of significance of the Fourier maps.

Selection of partially occupied solvent sites poses a most cumbersome problem in the modelling over and above that of the macromolecule itself, and can be highly subjective and very time consuming. Improved resolution of the data reveals additional weak or partially occupied solvent sites, which generally do not behave well during refinement. Water atoms modelled into relatively weak peaks in electron density tend to drift out of the density during refinement due to the inaccurate gradients that define their positions.

Given the huge number of water sites in question, automatic and at least semi-objective protocols are required. Several procedures have been developed for the automated identification of water sites during refinement [*inter alia ARP* (Lamzin & Wilson, 1997) and *SHELXL* (Sheldrick & Schneider, 1997)] and others allow selective inspection of such sites using graphics [such as *O* (Jones *et al.*, 1991) and *COOT* (Emsley & Cowtan, 2004)]. These depend on a combination of peak height in the density map and geometric considerations. However, these programs are currently optimized for structures at more typical resolutions, and future efforts could be made to adapt them for atomic resolution structures with overlapping water networks and other high levels of detail.

18.4.5.6. Bulk solvent and the low-resolution reflections

As stated in Section 18.4.5.3 and first reviewed by Matthews (1968) and more recently by Andersson & Hovmöller (1998), macromolecular crystals contain substantial regions of totally disordered, or bulk, aqueous solvent, in addition to those solvent molecules bound to the surface. The average electron density of the crystal volume occupied by protein is 1.35 g cm^{-3} (according to Matthews) or 1.22 g cm^{-3} (according to Andersson & Hovmöller), while that of water is 1.0 g cm^{-3}. This is because the atoms are more closely packed within the protein, as they are connected by covalent bonds, while in solvent regions they form sets of hydrogen-bonded networks.

To model both the solvent and protein regions of the crystal appropriately, it is necessary to have a satisfactory representation of the bulk solvent. The high *R* factors generally observed for most proteins for the low-resolution shells are in part symptomatic of the poor modelling of this feature or of systematic errors in the recording of the intensities of the low-angle reflections. For atomic resolution structures, the *R* factor can fall to values as low as 6–7% around 3–5 Å resolution. However, in lower-resolution shells it then rises steadily, often reaching values in the range of 20–40% below 10 Å. These observations indicate serious deficiencies in our current models or data.

The poorest approach is to ignore bulk solvent and assign zero electron density to those regions where there are no discrete atomic sites, as this leads to a severe discontinuum. An improved approach is to assign a constant value of the electron density to all points of the Fourier transform that are not covered by the discrete, ordered sites. This provides substantial reduction in the *R* factor for low-resolution shells of the order of 10% and requires the introduction of only one extra parameter to the least-squares minimization. An improvement of this simplistic model is the introduction of a second parameter, $B_{\rm sol}$, described by

$$\text{scale} = k_0 \exp(-B_0 s^2/4)[1 - k_{\rm sol} \exp(-B_{\rm sol} s^2/4)], \qquad (18.4.5.1)$$

where k_0 and B_0 are the scale factors for the protein, and $k_{\rm sol}$ and $B_{\rm sol}$ are the equivalent parameters for the bulk solvent (Tronrud, 1997). In effect, this provides a resolution-dependent smoothing of the interface contribution, rather than an overall term applied equally to all the data. The physical basis of this is discussed by Tronrud and implemented in several programs, for example *SHELXL*, *REFMAC* and *PHENIX.REFINE* (Fig. 18.4.5.2).

Another approach (Jiang & Brunger, 1994) to account for the effect of solvent is as follows: (1) calculate the mask covering the atoms in the crystal; (2) include a constant value for the electron density in the region not covered by the mask; (3) calculate structure factors from the solvent region; and (4) compute a total structure by applying the appropriate scale to the solvent contribution:

$$F_{\text{total}} = F_{\text{prot}} + k_m \exp(-B_m s^2/4) F_{\text{mask}}, \qquad (18.4.5.2)$$

where k_m and B_m are the scale and temperature factor for the solvent contribution, and F_{total}, F_{prot} and F_{mask} are the complex structure factors corresponding to the combined contribution from the protein and solvent region. The scale and B values are usually refined iteratively to find the best match to the observed structure factors, and are implemented in *CNS*, *REFMAC* and *PHENIX.REFINE*.

Nevertheless, there remain severe problems in the modelling of the interface. The border between the two regions is not abrupt, as there is a smooth and continuous change from the region with fully occupied, discrete sites to one which is truly fluid, but this passes through a volume with an increasing level of dynamic disorder and associated partial occupancy. Modelling of this region poses major problems, as described above, and the definition of disordered sites with low occupancy remains difficult even at atomic resolution. At which stage the occupancy and associated ADP can be defined with confidence is not yet an objective decision. In addition, refinement and modelling at this level of detail is very time consuming in terms of human intervention.

18.4.5.7. *Metal ions and other ligands in the solvent*

In general, proteins are crystallized from aqueous solutions which contain various additives, such as anions or cations (especially metals), organic solvents, including those used as cryoprotectants, and other ligands. Some of these may bind in specific or indeed non-specific sites in the ordered solvent shell, in addition to any functional binding sites of the protein. To identify such entities at limited resolution is often impossible, as the range of expected ADPs is large and there is very poor discrimination in the appearance of such sites and of water in the electron density. Atomic resolution assists in resolving ambiguities, as the interatomic distances, ADPs and occupancies are all better defined.

For metal ions, two additional criteria can be invoked. Firstly, the coordination geometry, with well defined bond lengths and angles, provides an indication of the identity of the ion, as different metals have different preferred ligand environments (Harding, 1999, 2006). The bond-valence approach is also applicable (Müller *et al.*, 2003). In addition, the value of the refined ADP and/or occupancy is helpful. Secondly, the anomalous signal in the data should reveal the presence of metal and some other non-water sites in the solvent through computation of the anomalous difference synthesis (Dauter & Dauter, 1999). This emphasizes the need to retain the anomalous signal during the collection and reduction of native data. While these approaches can be applied at lower resolution, they both become much more powerful at atomic resolution.

The presence of bound organic ligands has become especially relevant since the advent of cryogenic freezing. Compounds such as ethylene glycol and glycerol possess a number of functional hydrogen-bonding groups that can attach to sites on the protein in a defined way. Indeed, these may often bind in the active sites of enzymes such as glycosyl hydrolases, where they mimic the hydroxyl groups of the sugar substrate. It is most important to identify such moieties properly, particularly if substrate studies are to be planned successfully.

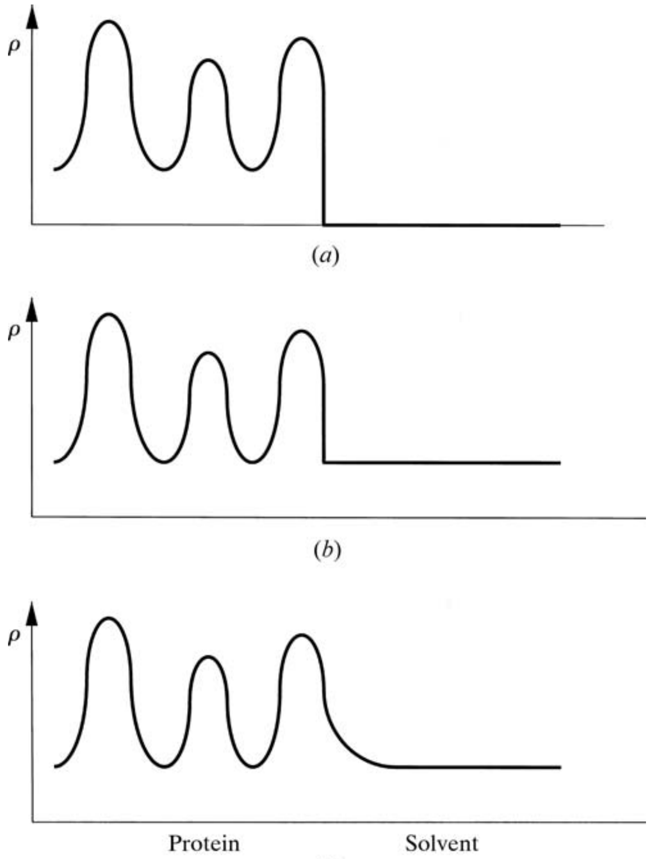

Figure 18.4.5.2
Schematic representation of the bulk-solvent models described in the text. (*a*) No bulk-solvent correction, *i.e.* solvent density set to zero. (*b*) Constant level of solvent outside the macromolecule and ordered water envelope. Here, sharp edge effects remain. (*c*) The model as in (*b*), but smoothed at the edge of a macromolecule, equivalent to the application of a *B* value to the solvent model.

18.4.5.8. *Deformation density*

X-ray structures are generally modelled using the spherical-atom approximation for the scattering, which ignores the deviation from sphericity of the outer bonding and lone-pair electrons. Extensive studies over a long period have confirmed that the so-called deformation density, representing deviation from this spherical model, can be determined experimentally using data to very high resolution, usually from 0.8 to 0.5 Å. An excellent review of this field has been provided by Coppens (1997). The observed deviations can be compared with those expected from the available theories of chemical bonding and the densities derived therefrom.

The application of atomic resolution analysis to proteins allowed the observation of the deformation density in macromolecules (Lamzin *et al.*, 1999). Data for two proteins were analysed: crambin (molecular weight 6 kDa) at 0.67 Å resolution and a subtilisin (molecular weight 30 kDa) at 0.9 Å resolution. Significant and interpretable deformation density could not be observed for the individual residues. However, on averaging the density over 40 peptide units for crambin and more than 250 for the subtilisin, the deformation density within the peptide unit was clearly visible and could be related to the expected bonding features in these units. This shows the real power of atomic resolution crystallography, which can reveal features containing no more than 0.2 e Å^{-3}.

Deformation density studies are now being applied to many polypeptides (Jelsch *et al.*, 1998; Koritsanszky *et al.*, 2002; Pichon-

Pesme *et al.*, 2004; Afonine *et al.*, 2007; Zarychta *et al.*, 2007). It has been observed that details of deformation densities are most clearly revealed when only the highest resolution (so-called high order) terms are included in the refinement and the Fourier maps (Coppens, 1997). This is reported to result from the deconvolution of the effects of the anisotropic ADPs, which can to some extent take up the fine features corresponding to the deformation density (bonding electrons and lone pairs). After proper modelling of the deformation density features, the overall model is refined against all the data.

18.4.6. Quality assessment of the model

The refinement of proteins at resolutions lower than atomic depends upon the use of restraints on the geometry and ADPs. The almost exclusively used library of target geometric restraints for refinement and validation of protein structures (Chapter 18.3) is derived from structures of amino acids and peptides in the Cambridge Crystallographic Data Centre's small-molecule crystal structure database (Allen *et al.*, 1979). Stereochemical parameters, such as conformational angles φ, ψ, should ideally not be restrained, as they allow independent validation of the model. As stated in Section 18.4.4.2, these restraints are required even at atomic resolution to maintain the chemical integrity of flexible regions, although their impact will be limited on ordered regions.

Owing to the excess of accurate X-ray observations over parameters at atomic resolution, extensive validation of individual structures should be less challenging than for those at lower resolution. It is hard to achieve an R factor around 10% with an incorrect model. However, considerable attention needs to be given to detail and great care taken to avoid over-interpretation, especially of the flexible regions.

An analysis of eight structures determined at atomic resolution some years ago (EU 3-D Validation Network, 1998) indicated that they follow the expected rules of chemistry more closely than those of lower-resolution analyses in the PDB, confirming that atomic resolution indeed provides more precise coordinates. A subsequent analysis of ten structures at ultra-high resolution, 0.8 Å or better (Jaskolski *et al.*, 2007) confirmed these conclusions but identified a few possible adjustments to some targets. Following this analysis, Karplus *et al.* (2008) proposed that protein stereochemistry is context dependent, *e.g.* it differs in detail between α-helices and β-strands, and that this should be reflected in future target libraries. Thus, the availability of atomic resolution structures will provide a more objective basis for the construction of such libraries.

18.4.7. Relation to biological chemistry

A question arises as to what biological issues are addressed by analysis of macromolecular structures at atomic resolution. For any protein, the overall structure of its fold, and hence its homology with other proteins, can already be provided by analyses at low to medium resolution. However, proteins are the active entities of cells and carry out recognition of other macromolecules, ligand binding and catalytic roles that depend upon subtle details of chemistry, for which accurate positioning of the atoms is required. Even at atomic resolution, the accuracy of structural definition is less than what would ideally be required for the changes observed during a chemical reaction. At lower resolutions, structure–function relations require yet further extrapolation of the experimental data.

To understand the function of many macromolecules, such as enzymes, it is not sufficient to determine the structure of a single state. Alongside the native structure, those of various complexes will also be required. The differences between the states provide additional information on the functionality. For an understanding of the chemistry involved, atomic resolution has tremendous advantages in terms of accuracy, as reliable judgments can be based on the experimental data alone.

Advantages of atomic resolution include the following:

(1) The positions of all atoms that possess defined conformations are more accurately defined. This means that all bond lengths and angles in the structure have lower standard uncertainties (EU 3-D Validation Network, 1998). For regions of the molecule where the conformation is representative of the norm, this is of purely quantitative significance, but where the stereochemistry deviates from the expected value this accuracy takes on a special significance, which poses questions to the theoretical chemist. Such deviations from standard geometry often play an important role in biological function.

(2) The better the ADP definition, notably its anisotropy, the greater the insight into the static or thermal flexibility of individual regions of the molecule. Macromolecules are crucially dependent upon flexibility for properties such as induced fit in substrate or ligand recognition, allosteric responses or responses to the biological environment. More detailed definition of the position and mobility of flexible regions may be assisted by atomic resolution analysis.

(3) A few amino-acid side chains play an active role in catalysis. Those that do include histidine, aspartic and glutamic acids and serine, all through protonation–deprotonation events, and hydrogen atoms are crucial to their function. Hydrogen atoms are usually treated as riding on their parent atoms and should be included in the model, even at medium resolution. Unfortunately, those hydrogen atoms that are of interest can only rarely be treated as rigidly bonded at a predictable position. However, atomic resolution allows many hydrogen atoms to be clearly identified in the refined electron density. In addition, the presence or absence of hydrogen may be inferred by accurate estimation of the bond lengths between atoms, *e.g.* within the carboxylate groups.

(4) The relative orientation of reacting moieties is crucial to enzyme catalysis. If chemical hypotheses of mechanism are to be subjected to appropriate Popperian scrutiny (Popper, 1959), then precise definition of atomic coordinates in native and complex structures is necessary.

(5) Enzyme catalysis provides a reduction of the activation energy of the reaction, which can be achieved by distortion of the conformation of the substrate bound to the enzyme (the so-called Michaelis complex) towards the transition state or by the stabilization of the latter by the enzyme. For both, the study of complexes of inhibitors or substrate analogues at a sufficient resolution to clarify the fine detail of the structures is required.

(6) Adaptation of the enzyme to the substrate is postulated by the induced-fit theory of catalysis. The level of adjustment can be very small, and energy calculations again require that this be precisely defined.

(7) In metalloproteins, the ligand field, and hence geometry and bond lengths, around the metal ion are essential indicators of any variation in valence electrons between different states. For example, bond lengths between oxidized and reduced states of metal ions vary by the order of 0.1 Å or less, and

clear distinction between alternative oxidation states requires an accuracy only provided by atomic resolution.

Almost all atomic resolution analyses require data recorded from cryogenically frozen crystals. This does pose some problems of biological relevance, as proteins *in vivo* have adapted to operate at ambient cellular temperatures. The structure that is required is that of the protein and surrounding solvent at the corresponding temperature. The trade-off is that cryogenic structures may be better defined, but that they are only so because of the increased order of the protein and solvent at low temperature. This has to be weighed against the lack of fine detail in a medium-resolution analysis at room temperature.

A question often raised with regard to the worth of atomic resolution data concerns the effort required in refining a protein at such resolution. To define all details, such as alternative conformations, hydrogen-atom positions and solvent, is certainly time consuming, especially if an anisotropic model is adopted. However, the advantages outweigh the disadvantages, as even if a full anisotropic model is not refined to exhaustion, nevertheless all density maps will be clearer if the resolution is better, resulting in an improved definition of the features of interest.

18.4.8. Practical strategies

Pioneering work was carried out by Teeter and colleagues on crambin, using data recorded on this small and highly stable protein using a conventional diffractometer (Teeter *et al.*, 1993). Some of the earliest atomic resolution structures using data from a synchrotron source with an imaging plate detector included rubredoxin at 1.0 Å (Dauter *et al.*, 1992), ribonuclease Sa at 1.1 Å (Sevcik *et al.*, 1996) and triclinic lysozyme at 0.9 Å resolution (Walsh *et al.*, 1998), at room temperature for the rubredoxin and ribonuclease, and 100 K for the lysozyme. The strategy used involved the application of conventional restrained least-squares or maximum-likelihood techniques in the early stages of refinement, followed by switching over to *SHELXL* to introduce a full anisotropic model and riding hydrogen atoms.

Subsequent structures have almost exclusively used cryogenically vitrified crystals to minimize the effects of radiation damage on high-intensity SR beamlines. A review of the field (Dauter, 2003) reported over 100 structures at better than 1.2 Å spacing, and in January 2011 there were around 1450 such coordinate sets in the PDB.

Initially, only *SHELXL* was able to refine macromolecular structures with anisotropic ADPs, riding hydrogen atoms and appropriate stereochemical restraints. Subsequently, such options have been implemented in programs such as *REFMAC* and *PHENIX.REFINE*. The use of fast Fourier algorithms gives these programs some speed advantage over *SHELXL*, but this is now of less importance given the power of contemporary computers. The key advantage of *SHELXL* is the ability to perform full- (or at least block-) matrix inversion, and hence extract error estimates for each individual parameter.

It is not straightforward to provide hard and fast rules for the optimum strategy. The power of refinement at atomic resolution lies in the sheer number of X-ray observations which drive the minimization to a global minimum. The dangers of bias in the final model and of false minima can largely be ignored. As a guideline the following steps would seem to be a common-sense approach when data to 1.2 Å or better are available.

(1) Refine an initial isotropic model at modest resolution with one of the conventional programs to ensure the correct global minimum has been reached. This step was probably more appropriate in the days when computing resources were more limiting, and could probably be omitted today.

(2) Introduce an anisotropic model with riding hydrogen atoms using one of the maximum-likelihood programs such as *REFMAC*, *PHENIX.REFINE* or *SHELXL* (with the conjugate-gradient option), and refine against the experimental data, omitting around 1000–2000 reflections for validation by $R_{\rm free}$. It is *not* necessary to omit 5% of the data, particularly at this resolution.

(3) Complete the water model using one of the autobuild options within or associated with (*e.g. ARP/wARP* or *COOT*) the refinement program. This is an area amenable to considerable automation in the future to allow partially occupied sites and overlapping networks to be defined without user intervention.

(4) Introduce alternative conformations for the protein. Much of this is also amenable to automation, especially for the side chains, where programs such as *COOT* can already suggest likely alternate rotamers.

(5) Inspect appropriately weighted electron-density maps for features to be added, deleted or moved. This includes ligands, metal atoms, ions, cryprotectants and other additives. Current developments in programs such as *COOT* are increasingly automating this step. Software libraries that provide accurate descriptions of the ligand chemistry and geometry are essential, and are the subject of intense development. It is vital that structures are deposited in the PDB with the correct geometrical restraints. To some extent, steps (3)–(5) go in parallel, and indeed their order is not vital.

(6) Consider the *SHELXL* full-matrix option for the final refinement cycles. This will provide estimates of the accuracy of the individual parameters, not available from other programs.

(7) As a final step, run some cycles with all reflections including those previously used for $R_{\rm free}$, but without changing the model parameters. At all stages, it will remain necessary to impose restraints to avoid poor behaviour of any flexible regions. As the resolution goes beyond 1 Å, the effect of such restraints on good parts of the structure becomes minimal.

In summary, at atomic resolution models require special attention to detail, reflecting the wealth of information with regard to features such as alternative conformations of both side and main chains, and extensive ordered water structure often with partial occupancy and overlapping networks. With the present software, this can be a lengthy operation, and indeed this can persuade some scientists to work at lower, less informative, resolution to speed up and simplify the analysis. Many of the presently tedious manual tasks in terms of model construction are, however, amenable to extensive automation. As the number of atomic resolution structures increases, we can expect the refinement and modelling process to be greatly simplified with new software algorithms.

References

Afonine, P. V., Grosse-Kunstleve, R. W., Adams, P. D., Lunin, V. Y. & Urzhumtsev, A. (2007). *On macromolecular refinement at subatomic resolution with interatomic scatterers*. *Acta Cryst.* D**63**, 1194–1197.

Agarwal, R. C. (1978). *A new least-squares refinement technique based on the fast Fourier transform algorithm*. *Acta Cryst.* A**34**, 791–809.

Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). *The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information*. *Acta Cryst.* B**35**, 2331–2339.

Andersson, K. M. & Hovmöller, S. (1998). *The average atomic volume and density of proteins*. *Z. Kristallogr.* **213**, 369–373.

Blessing, R. H. (1997). *LOCSCL: a program to statistically optimize local scaling of single-isomorphous-replacement and single-wavelength-anomalous-scattering data*. *J. Appl. Cryst.* **30**, 176–177.

Box, G. E. P. & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, California, London: Addison-Wesley.

Bricogne, G. & Irwin, J. J. (1996). *Maximum-likelihood structure refinement: theory and implementation within BUSTER+TNT*. In *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.

Brünger, A. T. (1992*a*). *Free R value: a novel statistical quantity for assessing the accuracy of crystal structures*. *Nature (London)*, **355**, 472–475.

Brünger, A. T. (1992*b*). *X-PLOR manual*. Version 3.1. New Haven: Yale University.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Crystallography & NMR system: a new software suite for macromolecular structure determination*. *Acta Cryst.* D**54**, 905–921.

Collaborative Computational Project, Number 4 (1994). *The CCP4 suite: programs for protein crystallography*. *Acta Cryst.* D**50**, 760–763.

Coppens, P. (1997). *X-ray Charge Densities and Chemical Bonding*. International Union of Crystallography and Oxford University Press.

Cowtan, K. D. & Main, P. (1998). *Miscellaneous algorithms for density modification*. *Acta Cryst.* D**53**, 487–493.

Cruickshank, D. W. J. (1999*a*). *Remarks about protein structure precision*. *Acta Cryst.* D**55**, 583–601.

Cruickshank, D. W. J. (1999*b*). *Remarks about protein structure precision. Erratum*. *Acta Cryst.* D**55**, 1108.

Dauter, Z. (2003). *Protein structures at atomic resolution*. *Methods Enzymol.* **368**, 288–337.

Dauter, Z. & Dauter, M. (1999). *Anomalous signal of solvent bromides used for phasing of lysozyme*. *J. Mol. Biol.* **289**, 93–101.

Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1997). *The benefits of atomic resolution*. *Curr. Opin. Struct. Biol.* **7**, 681–688.

Dauter, Z., Sieker, L. C. & Wilson, K. S. (1992). *Refinement of rubredoxin from Desulfovibrio vulgaris at 1.0 Å with and without restraints*. *Acta Cryst.* B**48**, 42–59.

Dauter, Z., Wilson, K. S., Sieker, L. C., Meyer, J. & Moulis, J.-M. (1997). *Atomic resolution (0.94 Å) structure of Clostridium acidurici ferredoxin. Detailed geometry of [4Fe-4S] clusters in a protein*. *Biochemistry*, **36**, 16065–16073.

Diamond, R. (1971). *A real-space refinement procedure for proteins*. *Acta Cryst.* A**27**, 436–452.

Driessen, H., Haneef, M. I. J., Harris, G. W., Howlin, B., Khan, G. & Moss, D. S. (1989). *RESTRAIN: restrained structure-factor least-squares refinement program for macromolecular structures*. *J. Appl. Cryst.* **22**, 510–516.

Emsley, P. & Cowtan, K. (2004). *Coot: model-building tools for molecular graphics*. *Acta Cryst.* D**60**, 2126–2132.

EU 3-D Validation Network (1998). *Who checks the checkers? Four validation tools applied to eight atomic resolution structures*. *J. Mol. Biol.* **276**, 417–436.

French, S. & Wilson, K. S. (1978). *On the treatment of negative intensity observations*. *Acta Cryst.* A**34**, 517–525.

Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-H., Olson, W. K. & Berman, H. M. (1996). *Geometric parameters in nucleic acids: sugar and phosphate constituents*. *J. Am. Chem. Soc.* **118**, 519–528.

Harding, M. M. (1999). *The geometry of metal–ligand interactions relevant to proteins*. *Acta Cryst.* D**55**, 1432–1443.

Harding, M. M. (2006). *Small revisions to predicted distances around metal sites in proteins*. *Acta Cryst.* D**62**, 678–682.

Herzberg, O. & Sussman, J. L. (1983). *Protein model building by the use of a constrained–restrained least-squares procedure*. *J. Appl. Cryst.* **16**, 144–150.

International Tables for Crystallography (2004). Vol. C. *Mathematical, Physical and Chemical Tables*, edited by E. Prince. Dordrecht: Kluwer Academic Publishers.

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 Cryst.* D**63**, 611–620.

Jelsch, C., Pichon-Pesme, V., Lecomte, C. & Aubry, A. (1998). *Transferability of multipole charge-density parameters: application to very high resolution oligopeptide and protein structures*. *Acta Cryst.* D**54**, 1306–1318.

Jiang, J. S. & Brunger, A. T. (1994). *Protein hydration observed by X-ray diffraction. Solvation properties of penicillopepsin and neuraminidase crystal structures*. *J. Mol. Biol.* **243**, 100–115.

Johnson, C. K. (1976). *ORTEPII. A FORTRAN thermal-ellipsoid plot program for crystal structure illustration*. Report ORNL-5138. Oak Ridge National Laboratory, Tennessee, USA.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Improved methods for building protein models in electron density maps and the location of errors in these models*. *Acta Cryst.* A**47**, 110–119.

Karplus, P. A., Shapovalov, M. V., Dunbrack, R. L. Jr & Berkholz, D. S. (2008). *A forward-looking suggestion for resolving the stereochemical restraints debate: ideal geometry functions*. *Acta Cryst.* D**64**, 335–336.

Koritsanszky, T., Volkov, A. & Coppens, P. (2002). *Aspherical-atom scattering factors from molecular wave functions. 1. Transferability and conformation dependence of atomic electron densities of peptides within the multipole formalism*. *Acta Cryst.* A**58**, 464–472.

Konnert, J. H. & Hendrickson, W. A. (1980). *A restrained-parameter thermal-factor refinement procedure*. *Acta Cryst.* A**36**, 344–350.

Lamzin, V. S., Morris, R. J., Dauter, Z., Wilson, K. S. & Teeter, M. M. (1999). *Experimental observation of bonding electrons in proteins*. *J. Biol. Chem.* **274**, 20753–20755.

Lamzin, V. S. & Wilson, K. S. (1997). *Automated refinement for protein crystallography*. *Methods Enzymol.* **277**, 269–305.

Lebedev, A. A., Vagin, A. A. & Mushudov, G. N. (2006). *Intensity statistics in twinned crystals with examples from the PDB*. *Acta Cryst.* D**62**, 83–95.

Matthews, B. W. (1968). *Solvent content in protein crystals*. *J. Mol. Biol.* **33**, 491–497.

Morris, R. J., Blanc, E. & Bricogne, G. (2004). *On the interpretation and use of $<|E|^2>(d^*)$ profiles*. *Acta Cryst.* D**60**, 227–240.

Morris, R. J. & Bricogne, G. (2003). *Sheldrick's 1.2 Å rule and beyond*. *Acta Cryst.* D**59**, 615–617.

Müller, P., Köpke, S. & Sheldrick, G. M. (2003). *Is the bond-valence method able to identify metal atoms in protein structures?* *Acta Cryst.* D**59**, 32–37.

Murshudov, G. N., Davies, G. J., Isupov, M., Krzywda, S. & Dodson, E. J. (1998). *The effect of overall anisotropic scaling in macromolecular refinement*. In *CCP4 Newsletter on Protein Crystallography*, **35**, 37–42.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Refinement of macromolecular structures by the maximum-likelihood method*. *Acta Cryst.* D**53**, 240–255.

Murshudov, G. N., Vagin, A. A., Lebedev, A., Wilson, K. S. & Dodson, E. J. (1999). *Efficient anisotropic refinement of macromolecular structures using FFT*. *Acta Cryst.* D**55**, 247–255.

O'Hagan, A. (1994). *Kendal's Advanced Theory of Statistics; Bayesian Inference*, Vol. 2B. Cambridge: Arnold/Hodder Headline/Cambridge University Press.

Painter, J. & Merritt, E. A. (2006). *Optimal description of a protein structure in terms of multiple groups undergoing TLS motion*. *Acta Cryst.* D**62**, 439–450.

Pannu, N. S. & Read, R. J. (1996). *Improved structure refinement through maximum likelihood*. *Acta Cryst.* A**52**, 659–668.

Pichon-Pesme, V., Jelsch, C., Guillot, B. & Lecomte, C. (2004). *A comparison between experimental and theoretical aspherical-atom scattering factors for charge-density refinement of large molecules*. *Acta Cryst.* A**60**, 204–208.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

Prince, E. & Boggs, P. T. (2004). *International Tables for Crystallography*, Vol. C, ch. 8.1. Dordrecht: Kluwer Academic Publishers.

Schomaker, V. & Trueblood, K. N. (1968). *On the rigid-body motion of molecules in crystals*. *Acta Cryst.* B**24**, 63–76.

Schwarzenbach, D., Abrahams, S. C., Flack, H. D., Prince, E. & Wilson, A. J. C. (1995). *Statistical descriptors in crystallography. II. Report of a working group on expression of uncertainty in measurement*. *Acta Cryst.* A**51**, 565–569.

Sevcik, J., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1996). *Ribonuclease from Streptomyces aureofaciens at atomic resolution. Acta Cryst.* D**52**, 327–344.

Sheldrick, G. M. (1990). *Phase annealing in SHELX-90: direct methods for larger structures. Acta Cryst.* A**46**, 467–473.

Sheldrick, G. M. (2008). *A short history of SHELX. Acta Cryst.* A**64**, 112–122.

Sheldrick, G. M. & Schneider, T. R. (1997). *SHELXL: high-resolution refinement. Methods Enzymol.* **277**, 319–343.

Sheriff, S. & Hendrickson, W. A. (1987). *Description of overall anisotropy in diffraction from macromolecular crystals. Acta Cryst.* A**43**, 118–121.

Steiner, R. A., Lebedev, A. A. & Murshudov, G. N. (2003). *Fisher's information in maximum-likelihood macromolecular crystallographic refinement. Acta Cryst.* D**59**, 2114–2124.

Strokopytov, B. V. (2008). *How to multiply a matrix of normal equations by an arbitrary vector using FFT. Acta Cryst.* A**64**, 601–612.

Stuart, A., Ord, K. J. & Arnold, S. (1999). *Kendall's Advanced Theory of Statistics; Classical Inference and Linear Model*, Vol. 2A. London, Sydney, Auckland: Arnold/Hodder Headline.

Teeter, M. M., Roe, S. M. & Heo, N. H. (1993). *Atomic resolution (0.83 Å) crystal structure of the hydrophobic protein crambin at 130 K. J. Mol. Biol.* **230**, 292–311.

Ten Eyck, L. F. (1973). *Crystallographic fast Fourier transforms. Acta Cryst.* A**29**, 183–191.

Ten Eyck, L. F. (1977). *Efficient structure-factor calculation for large molecules by the fast Fourier transform. Acta Cryst.* A**33**, 486–492.

Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Zwart, P. H., Hung, L.-W., Read, R. J. & Adams, P. D. (2008). *Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard*. *Acta Cryst*. D**64**, 61–69.

Tickle, I. J. (2007). *Experimental determination of optimal root-mean-square deviations of macromolecular bond lengths and angles from their restrained ideal values*. *Acta Cryst.* D**63**, 1274–1281.

Tronrud, D. E. (1997). *TNT refinement package. Methods Enzymol.* **277**, 243–268.

Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Cryst*. D**55**, 191–205.

Walsh, M. A., Schneider, T. R., Sieker, L. C., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1998). *Refinement of triclinic hen egg-white lysozyme at atomic resolution. Acta Cryst.* D**54**, 522–546.

Wilson, A. J. C. (1942). *Determination of absolute from relative X-ray data intensities. Nature (London)*, **150**, 151–152.

Winn, M. D., Isupov, M. N. & Murshudov, G. N. (2001). *Use of TLS parameters to model anisotropic displacements in macromolecular refinement*. *Acta Cryst.* D**57**, 122–133.

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

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

Zarychta, B., Pichon-Pesme, V., Guillot, B., Lecomte, C. & Jelsch, C. (2007). *On the application of an experimental multipolar pseudo-atom library for accurate refinement of small-molecule and protein crystal structures*. *Acta Cryst.* A**63**, 108–125.