International Tables for Crystallography (2006). Vol. F, Chapter 18.4, pp. 393–402.

18.4. Refinement at atomic resolution

BY Z. DAUTER, G. N. MURSHUDOV AND K. S. WILSON

18.4.1. Definition of atomic resolution

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 can be 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 eleven parameters (Table 18.4.1.1).

The first parameter is the scattering-factor amplitude for the chemical nature of the atom in question, computed and tabulated for all atom types [*International Tables for Crystallography*, Volume C (1999)]. 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.

At atomic resolution, six anisotropic atomic displacement parameters are used to describe the distribution of the atoms 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 generally 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), $U_{\rm torsion}$ 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 twenty per independent moiety for $U_{\rm TLS}$. In contrast, refinement of the third contributor does pose a problem, as there is a 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 effect is indistinguishable.

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 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 chemistry of the atom (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 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.

Thus, at atomic resolution, minimization of the discrepancy between the experimentally determined amplitudes or intensities of the Bragg reflections and those calculated from the atomic model requires refinement of, at most, ten (usually nine) independent parameters per atom. This has been achieved classically by least squares, as described in *IT* C (1999), or more recently by maximum-likelihood procedures (Bricogne & Irwin, 1996; Pannu & Read, 1996; Murshudov *et al.*, 1997).

Atomicity is the great simplifying feature of crystallography in terms of structure solution and refinement. If atomic resolution is achieved, there are sufficient accurately measured observables to refine a full atomic model for the ordered part of the structure, but this condition can only be defined somewhat subjectively. A

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

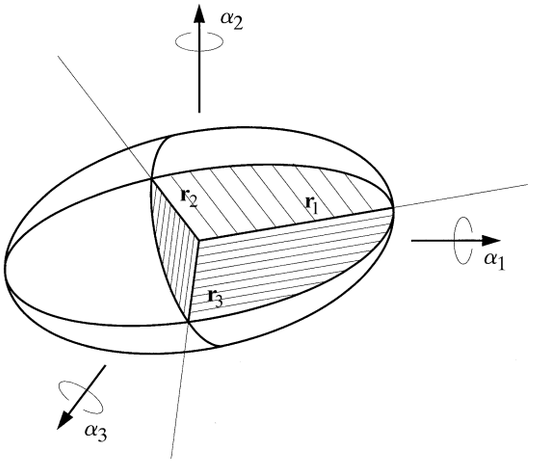

Fig. 18.4.1.1. The thermal-ellipsoid model used to represent anisotropic atomic displacement, with major axes indicated. The ellipsoid is drawn with a specified probability of finding an atom inside its contour. Six parameters are necessary to describe the ellipsoid: three represent the dimensions of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

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
1.5	Hydrogen atoms, 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

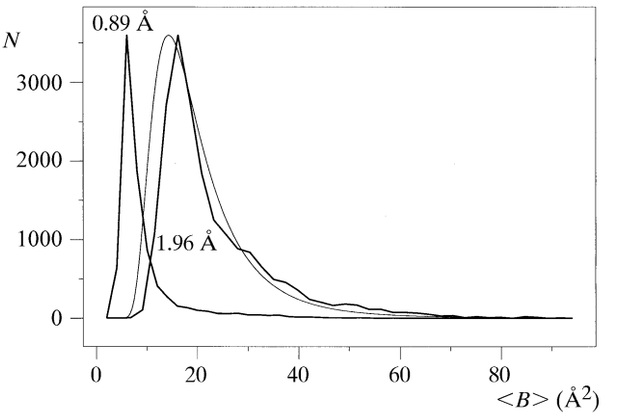

Fig. 18.4.1.2. Histograms of B values for a protein structure, *Micrococcus lysodecticus* catalase (Murshudov *et al.*, 1999), for two different crystals which diffracted to different limiting resolutions. For both crystals, the resolution cutoff reflects the real diffraction limit from the sample, and 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 is 25.5 Å^2 and the spread 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.

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). In practice, this means the statistical problem of refinement is overdetermined. For small-molecule structures, accurate amplitude data are normally available to around 0.8 Å, giving an observation-to-parameter ratio of about seven, allowing positional parameters to be determined with an accuracy of around 0.001 Å. This reflects the high degree of order of such crystals, in which the molecules in the lattice are in a closely packed array.

Crystals of macromolecules deviate substantially from this ideal. 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 may lead to a resolution cutoff much less than atomic. Thirdly, the large solvent content leads to substantial decay of crystal quality under exposure to the X-ray beam, especially at room temperature. 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).

As expected, atoms with different ADPs contribute differently to the diffraction intensities, as discussed by Cruickshank (1999). 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, 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 and macromolecular crystallography, but 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. However, it is now tractable to refine a significant number of proteins at atomic resolution with a full anisotropic model (Dauter, Lamzin & Wilson, 1997). This work of course benefits tremendously from the experience and algorithms of small-molecule crystallography, but it does pose special problems of its own. The techniques 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 & Schneider, 1997), which was developed for small structures and has now been extended to treat macromolecules.

An alternative and probably better approach to the definition of atomic resolution would be to employ 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. The appropriate choice of A allows the estimation of information on selected key features, such as the active site.

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) separate states is needed, the prior knowledge can be considered instead as knowledge about the native protein.

However, 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 gives some idea about 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 crystal forms is to reduce the limit of data necessary for achieving atomic resolution, which is equivalent to the advantage of noncrystallographic averaging.

18.4.1.1. *Ab initio phasing and atomic resolution*

Ab initio methods of phase calculation normally depend on the assumption of positivity and atomicity of the electron density. Such methods rely largely on the availability of atomic resolution data. In addition, approaches such as solvent flattening and automated map interpretation benefit enormously from such data. 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 atomic resolution.

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 synchrotron radiation (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 provide a saving in terms of the time spent by the investigator, since the interpretation of the resulting electron density is much easier and faster. In general, to answer the same question is much easier and cheaper if high-resolution data are available. In addition, high-resolution data mean that answers to some of the questions which may arise during analysis of the experiment will already be addressable. In contrast, low-resolution data not only make it difficult to answer the question currently being asked, but may 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. The rejection of low-intensity observations will have a similar biasing effect. 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), removes the overall crystal contribution to the ADP.

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 *IT* C (1999). Least squares involves the construction of a matrix of order $N \times N$, 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 standard uncertainties for each individual parameter refined. 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 that 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. 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.

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 uncertainty and from a more limited range and speed of convergence.

The first approach is the block-matrix approximation, 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) and *CNS* (Brünger *et al.*, 1998), but have been largely restricted to the diagonal approximation. *XPLOR* and *CNS* use the conjugate-gradient method that relies only on the first derivatives, ignoring 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 more statistically sound 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 (Murshudov *et al.*, 1997; Bricogne, 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 near future.

18.4.3.5. *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, which contain 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, which makes it hard to predict the optimal strategies for atomic resolution refinement, even over the next ten years.

18.4.4. Computational options and tactics

18.4.4.1. *Use of F or F^2*

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 can 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. For intensity-based likelihood, this is less important 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 assumed Wilson distribution could be upgraded using the scaling techniques suggested by Cowtan & Main (1998) and Blessing (1997), 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 $F_{\rm calc}$ when $F_{\rm calc}$ is equal to zero is singular (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 F^2, Fourier syntheses always require structure-factor amplitudes.

18.4.4.2. *Restraints and/or constraints on coordinates and ADPs*

Even for small-molecule structures, disordered regions of the unit cell require the imposition of stereochemical restraints or constraints if the chemical integrity is to be preserved and the ADPs are to be realistic. The restraints are comparable to those used for proteins at lower resolution and this makes sense, since the poorly ordered regions with high ADPs in effect do not contribute to the high-angle diffraction terms, and as a result their parameters are only defined by the lower-angle amplitudes.

Thus, even for a macromolecule for which the crystals diffract to atomic resolution, there will be regions possessing substantial thermal or static disorder, and restraints on the positional parameters and ADPs are essential for these parts. Their effect on the ordered regions will be minimal, as the X-ray terms will dominate the refinement, provided the relative weighting of X-ray and geometric contributions is appropriate.

Another justification for use of restraints is that refinement can be considered a 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, and secondly, different geometric parameters are correlated. This problem should be the subject of further investigation.

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 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. Early attempts to derive such an objective tool were made by Hamilton (1965). Unfortunately, they proved to be cumbersome in practice for large structures and did not provide the required objectivity.

Direct application of the Hamilton test is especially problematical for macromolecules because of the use of restraints. Attempts have been made to overcome these problems, using a direct extension of the Hamilton test itself (Bacchi *et al.*, 1996) or with a combination of self and cross validation (Tickle *et al.*, 1998).

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 presently being developed, 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 reduction in the number of atoms, but this is not an ideal procedure, as real chemical entities of the model are sacrificed. Reducing the number of 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 not all water molecules cannot 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.

18.4.4.5. *Practical strategies*

It is not reasonable to give absolute rules for refinement of atomic resolution structures at this time, as the field is rather new and is developing rapidly. Pioneering work has been carried out by Teeter *et al.* (1993) on crambin, based on data recorded on this small and highly stable protein using a conventional diffractometer. Studies on perhaps more representative proteins are those on ribonuclease Sa at 1.1 Å (Sevcik *et al.*, 1996) and triclinic lysozyme at 0.9 Å resolution (Walsh *et al.*, 1998). These studies used data from a synchrotron source with an imaging-plate detector at room temperature for the ribonuclease and at 100 K for the lysozyme. The strategy involved the application of conventional restrained least squares or maximum-likelihood techniques in the early stages of refinement, followed by a switch over to *SHELXL* to introduce a full anisotropic model. A series of other papers have appeared in the literature following similar protocols, reflecting the fact that, until recently, only *SHELXL* was generally available for refining macromolecular structures with anisotropic models and appropriate stereochemical restraints. Programs such as *REFMAC* have now been extended to allow anisotropic models. As they use fast Fourier transforms for the structure-factor calculations, the speed advantage will mean that *REFMAC* or comparable programs are likely to be used extensively in this area in the future, even if *SHELXL* is used in the final step to extract error estimates.

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

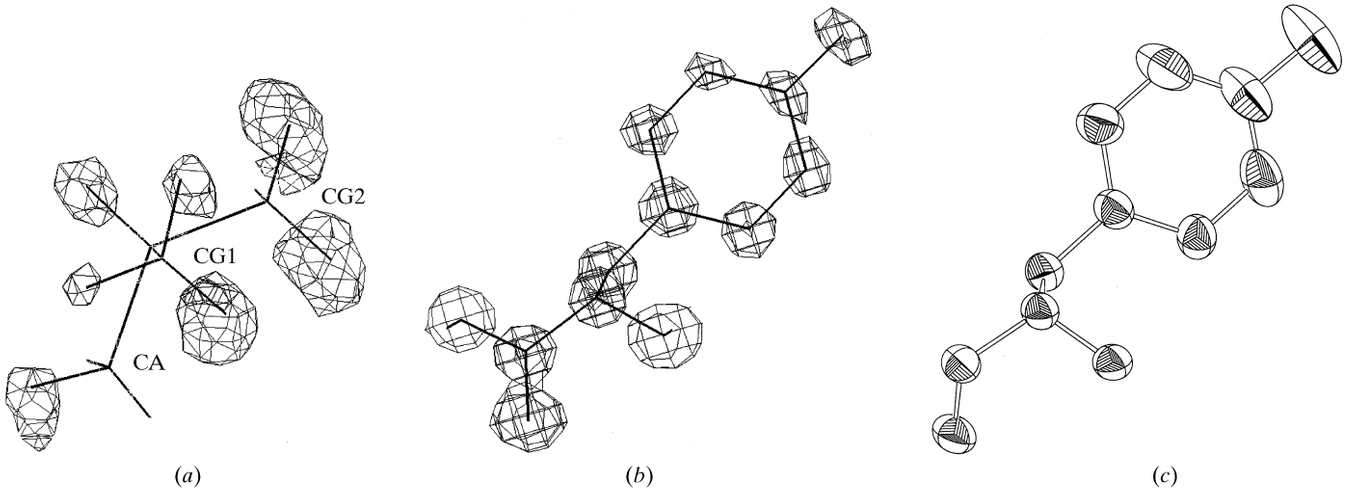

Fig. 18.4.5.1. (*a*), (*b*) Representative electron-density maps for the refinement of *Clostridium acidurici* ferredoxin at 0.94 Å resolution (Dauter, Wilson *et al.*, 1997). (*a*) The density for hydrogen atoms (at 3σ) omitted from the structure-factor calculation for Val42. (*b*) The $(2F_o - F_c)$ 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.

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 and anisotropic parameters (see below) 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 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* (Sheldrick & Schneider, 1997) or *HGEN* from the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994). 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.

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, 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 recently, anisotropic ADPs have 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).

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 below).

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, be they side- or main-chain, 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 some 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 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 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 means 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 weak 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 [*O* (Jones *et al.*, 1991) and *Quanta* (Molecular Simulations Inc., San Diego)]. These depend on a combination of peak height in the density map and geometric considerations.

18.4.5.6. *Bulk solvent and the low-resolution reflections*

As stated in the preceding section 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 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 partly 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

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

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 data. The physical basis of this is discussed by Tronrud and implemented in several programs, for example *SHELXL* (Sheldrick & Schneider, 1997) and *REFMAC* (Murshudov *et al.*, 1997) (Fig. 18.4.5.2).

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

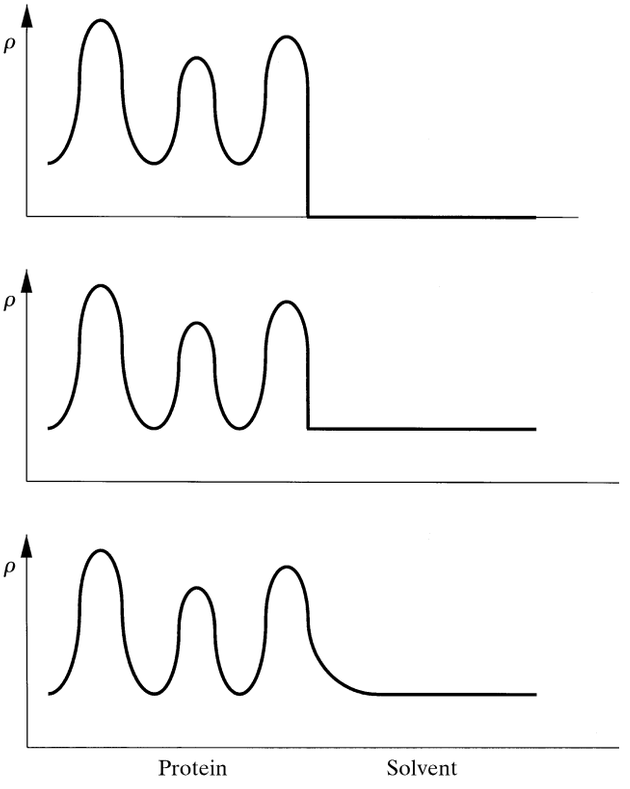

Fig. 18.4.5.2. Schematic representation of the bulk-solvent models described in the text. (*a*) No bulk-solvent correction, *i.e.* solvent density set to zero. (*b*) Constant level of solvent outside the macromolecule and ordered water envelope. Here, sharp edge effects remain. (*c*) The model as in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

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 all the interatomic distances, ADPs and occupancies are 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 [see, for example, Nayal & Di Cera (1996)]. 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 by computation of the anomalous difference synthesis (Dauter & Dauter, 1999). 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.

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 recent review of this field is 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. Such studies have been applied to peptides and related molecules (Souhassou *et al.*, 1992; Jelsch *et al.*, 1998).

The application of atomic resolution analysis to proteins has allowed the first steps towards 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. 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}.

18.4.6. Quality assessment of the model

The refinement of proteins at resolution lower than atomic depends upon the use of restraints on the geometry and ADPs. Most target libraries for refinement and validation of structures (*e.g.* Engh & Huber, 1991) are derived from either the Cambridge Structural Database (Allen *et al.*, 1979) or from protein structures in the Protein Data Bank (PDB; Bernstein *et al.*, 1977). The availability of atomic resolution structures provides more objective data for the construction of target libraries. Stereochemical parameters, such as conformational angles φ, ψ, should ideally not be restrained, as they allow independent validation of the model. Analysis of eight structures determined at atomic resolution (EU 3-D Validation Network, 1998) indicates 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.

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 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) throughout 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 required structure is that of the protein and surrounding solvent at the corresponding temperature. The trade-off is that cryogenic structures may be better defined, but only because of the increased order of 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.

18.2 (*cont.*)

1.8 Å resolution of the aspartic proteinase from Rhizopus chinensis. J. Mol. Biol. **196**, 877–900.
Verlet, L. (1967). *Computer experiments on classical fluids. I. Thermodynamical properties of Lennard–Jones molecules. Phys. Rev*. **159**, 98–105.

18.3

Bansal, M. & Ananthanarayanan, V. S. (1988). *The role of hydroxyproline in collagen folding: conformational energy calculations on oligopeptides containing proline and hydroxyproline. Biopolymers*, **27**, 299–312.
Bricogne, G. (1993). *Direct phase determination by entropy maximization and likelihood ranking: status report and perspectives. Acta Cryst.* D**49**, 37–60.
Brünger, A. T. (1993). *Assessment of phase accuracy by cross validation: the free R value. Methods and applications. Acta Cryst.* D**49**, 24–36.
Bürgi, H.-B. & Dubler-Steudle, K. C. (1988). *Empirical potential energy surfaces relating structure and activation energy. 2. Determination of transition-state structure for the spontaneous hydrolysis of axial tetrahydropyranyl acetals. J. Am. Chem. Soc.* **110**, 7291–7299.
Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1997). *The benefits of atomic resolution. Curr. Opin. Struct. Biol.* **7**, 681–688.
Engh, R. A. & Huber, R. (1991). *Accurate bond and angle parameters for X-ray protein structure refinement. Acta Cryst.* A**47**, 392–400.
Hooft, R. W. W., Sander, C. & Vriend, G. (1997). *Objectively judging the quality of a protein structure from a Ramachandran plot. Comput. Appl. Biosci.* **13**, 425–430.
Kidera, A., Matsushima, M. & Go, N. (1994). *Dynamic structure of human lysozyme derived from X-ray crystallography: normal mode refinement. Biophys. Chem.* **50**, 25–31.
Kleywegt, G. J. & Jones, T. A. (1998). *Databases in protein crystallography. Acta Cryst.* D**54**, 1119–1131.
Lamzin, V. S., Dauter, Z. & Wilson, K. S. (1995). *Dictionary of protein stereochemistry. J. Appl. Cryst.* **28**, 338–340.
Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst.* **26**, 283–291.
Laskowski, R. A., Moss, D. S. & Thornton, J. M. (1993). *Main-chain bond lengths and bond angles in protein structures. J. Mol. Biol.* **231**, 1049–1067.
Longhi, S., Czjzek, M. & Cambillau, C. (1998). *Messages from ultrahigh resolution crystal structures. Curr. Opin. Struct. Biol.* **8**, 730–737.
Marquart, M., Walter, J., Deisenhofer, J., Bode, W. & Huber, R. (1983). *The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. Acta Cryst.* B**39**, 480–490.
Pannu, N. S. & Read, R. J. (1996). *Improved structure refinement through maximum likelihood. Acta Cryst.* A**52**, 659–668.
Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A. T. & Berman, H. M. (1996). *New parameters for the refinement of nucleic acid-containing structures. Acta Cryst.* D**52**, 57–64.
Priestle, J. P. (1994). *Stereochemical dictionaries for protein structure refinement and model building. Structure*, **2**, 911–913.
Sippl, M. J. (1995). *Knowledge-based potentials for proteins. Curr. Opin. Struct. Biol.* **5**, 229–235.
Stewart, D. E., Sarkar, A. & Wampler, J. E. (1990). *Occurrence and role of cis peptide bonds in protein structures. J. Mol. Biol.* **214**, 253–260.
Vlassi, M., Dauter, Z., Wilson, K. S. & Kokkinidis, M. (1998). *Structural parameters for proteins derived from the atomic resolution (1.09 Å) structure of a designed variant of the ColE1 ROP protein. Acta Cryst.* D**54**, 1245–1260.
Wilson, K. S., Butterworth, S., Dauter, Z., Lamzin, V. S., Walsh, M., Wodak, S., Pontius, J., Richelle, J., Vaguine, A., Sander, C., Hooft, R. W. W., Vriend, G., Thornton, J. M., Laskowski, R. A., MacArthur, M. W., Dodson, E. J., Murshudov, G., Oldfield, T. J., Kaptein, R. & Rullmann, J. A. C. (1998). *Who checks the checkers – four validation tools applied to eight atomic resolution structures. J. Mol. Biol.* **276**, 417–436.

18.4

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.
Bacchi, A., Lamzin, V. S. & Wilson, K. S. (1996). *A self-validation technique for protein structure refinement: the extended Hamilton test. Acta Cryst.* D**52**, 641–646.
Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. E., Brice, M. D., Rogers, J. K., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol.* **112**, 535–542.
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, Massachusetts/California/London: Addison-Wesley.
Bricogne, G. (1997). *Maximum entropy methods and the Bayesian programme*. In *Proceedings of the CCP4 study weekend. Recent advances in phasing*, edited by K. S. Wilson, G. Davies, A. W. Ashton & S. Bailey, pp. 159–178. Warrington: Daresbury Laboratory.
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). *Remarks about protein structure precision. Acta Cryst.* D**55**, 583–601; erratum (1999), D**55**, 1108.
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., 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.

18.4 (*cont.*)

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.

Engh, R. A. & Huber, R. (1991). *Accurate bond and angle parameters for X-ray protein structure refinement*. *Acta Cryst.* A**47**, 392–400.

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.

Hamilton, W. C. (1965). *Significance tests on the crystallographic R factor. Acta Cryst.* **18**, 502–510.

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 (1999). Vol. C. *Mathematical, physical and chemical tables*, edited by A. J. C. Wilson & E. Prince. Dordrecht: Kluwer Academic Publishers.

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.

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.

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.

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

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.

Nayal, M. & Di Cera, E. (1996). *Valence screening of water in protein crystals reveals potential Na^+ binding sites. J. Mol. Biol.* **256**, 228–234.

O'Hagan, A. (1994). *Kendal's advanced theory of statistics; Bayesian inference*, Vol. 2B. Cambridge: Arnold, Hodder Headline and Cambridge University Press.

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

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.

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. & 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.

Souhassou, M., Lecomte, C., Ghermani, N.-E., Rohmer, M.-M., Roland, W., Benard, M. & Blessing, R. H. (1992). *Electron distributions in peptides and related molecules. 2. An experimental and theoretical study of (Z)-N-acetyl-α,β-dehydrophenylalanine methylamide. J. Am. Chem. Soc.* **114**, 2371–2382.

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.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998). *R_{free} and the R_{free} ratio. Part I: Derivation of expected values of cross-validation residuals used in macromolecular least-squares refinement. Acta Cryst.* D**54**, 547–557.

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

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.

18.5

Allen, F. H., Cole, J. C. & Howard, J. A. K. (1995*a*). *A systematic study of coordinate precision in X-ray structure analyses. I. Descriptive statistics and predictive estimates of e.s.d.'s for C atoms. Acta Cryst.* A**51**, 95–111.

Allen, F. H., Cole, J. C. & Howard, J. A. K. (1995*b*). *A systematic study of coordinate precision in X-ray structure analyses. II. Predictive estimates of e.s.d.'s for the general-atom case. Acta Cryst.* A**51**, 112–121.

Bricogne, G. (1993*a*). *Direct phase determination by entropy maximization and likelihood ranking: status report and perspectives. Acta Cryst.* D**49**, 37–60.

Bricogne, G. (1993*b*). *Fourier transforms in crystallography: theory, algorithms, and applications*. In *International tables for crystallography*, Vol. B, edited by U. Shmueli, pp. 88–89. Dordrecht: Kluwer Academic Publishers.

Bricogne, G. & Irwin, 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). *Free R-value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature (London)*, **355**, 472–475.

Brünger, A. T. (1993). *Assessment of phase accuracy by cross validation: the free R value. Methods and application. Acta Cryst.* D**49**, 24–36.

Chambers, J. L. & Stroud, R. M. (1979). *The accuracy of refined protein structures: comparison of two independently refined models of bovine trypsin. Acta Cryst.* B**35**, 1861–1874.

Cochran, W. (1948). *The Fourier method of crystal-structure analysis. Acta Cryst.* **1**, 138–142.