International Tables for Crystallography (2006). Vol. F, Section 18.4.4, pp. 396–398.

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

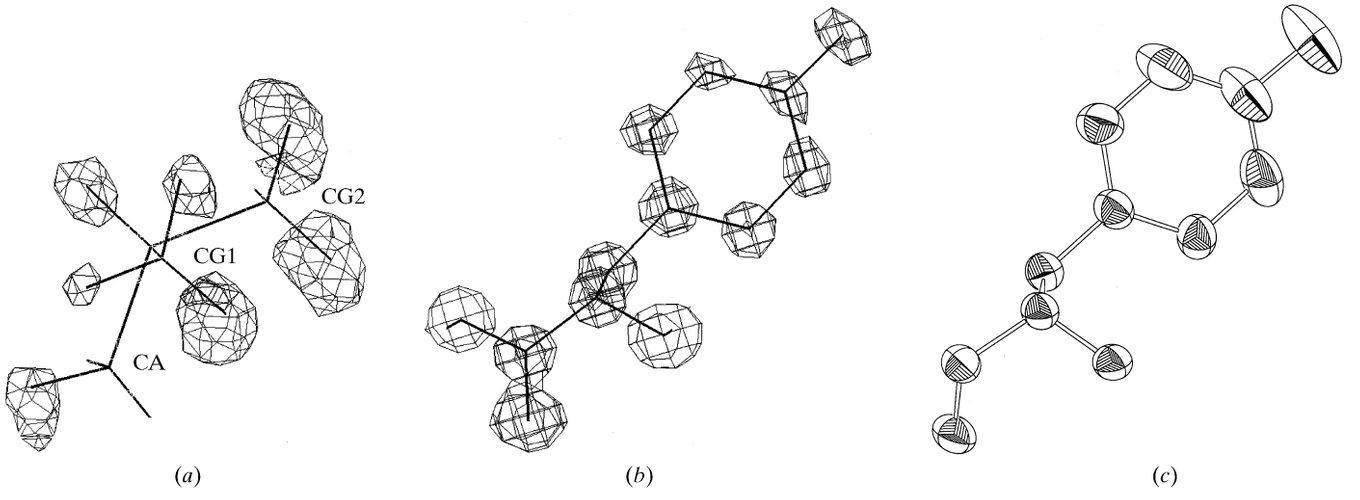

Fig. 18.4.5.1. (*a*), (*b*) Representative electron-density maps for the refinement of *Clostridium acidurici* ferredoxin at 0.94 Å resolution (Dauter, Wilson *et al.*, 1997). (*a*) The density for hydrogen atoms (at 3σ) omitted from the structure-factor calculation for Val42. (*b*) The $(2F_o - F_c)$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

references