gradient approach, all diagonal terms of the matrix are equal. However, the range and speed of convergence are substantially reduced, and standard uncertainties can no longer be estimated directly by matrix inversion.

18.4.3.3. Fast Fourier transform

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

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

18.4.3.4. Maximum likelihood

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

18.4.3.5. Twinning

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

18.4.3.6. Computer power

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

18.4.4. Computational options and tactics

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

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

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

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

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

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

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

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

18.4.4.2. Restraints on coordinates and ADPs

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

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

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

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

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

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

18.4.4.3. Partial occupancy

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

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

(1) when a protein or ligand atom is present in all molecules in the lattice, but can lie in more than one position due to alternative conformations;
(2) for the solvent region, where there may be overlapping and mutually exclusive solvent networks;
(3) when ligand-binding sites are only partially occupied due to weak binding constants, and the structures represent a mixture of native enzyme with associated solvent and the complex structure;
(4) when there is a mixture of protein residues in the crystal, due to inhomogeneity of the sample arising from polymorphism, a mixture of mutant and wild-type protein, or other causes.

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

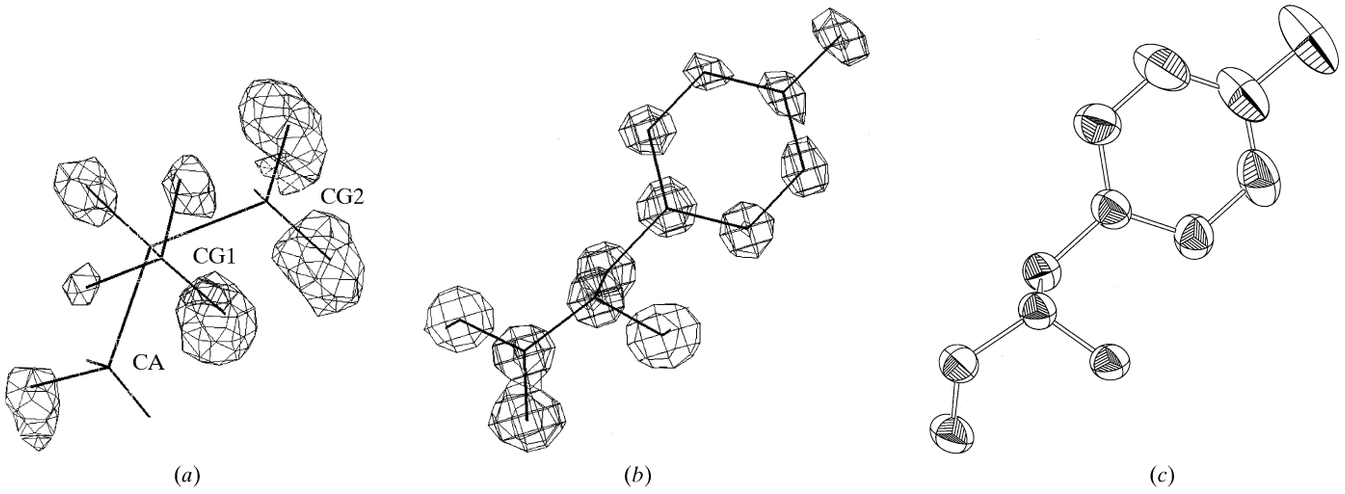

Figure 18.4.5.1
(*a*), (*b*) Representative electron-density maps for the refinement of *Clostridium acidurici* ferredoxin at 0.94 Å resolution (Dauter, Wilson *et al.*, 1997). (*a*) The density for hydrogen atoms (at 3σ) omitted from the structure-factor calculation for Val42. (*b*) The $(2F_o - F_c)$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

prove difficult to refine the occupancy satisfactorily with statistical certainty.

18.4.4.4. Validation of extra parameters during the refinement process

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

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

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

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

18.4.5. Features in the refined model

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

18.4.5.1. Hydrogen atoms

Hydrogen atoms possess only a single electron and therefore have low electron density and are relatively poorly defined in X-ray studies. They play central roles in the function of proteins,