

## 18. REFINEMENT

limitations to this. Firstly, if  $R_{\text{free}}$  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  $R_{\text{free}}$ , 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

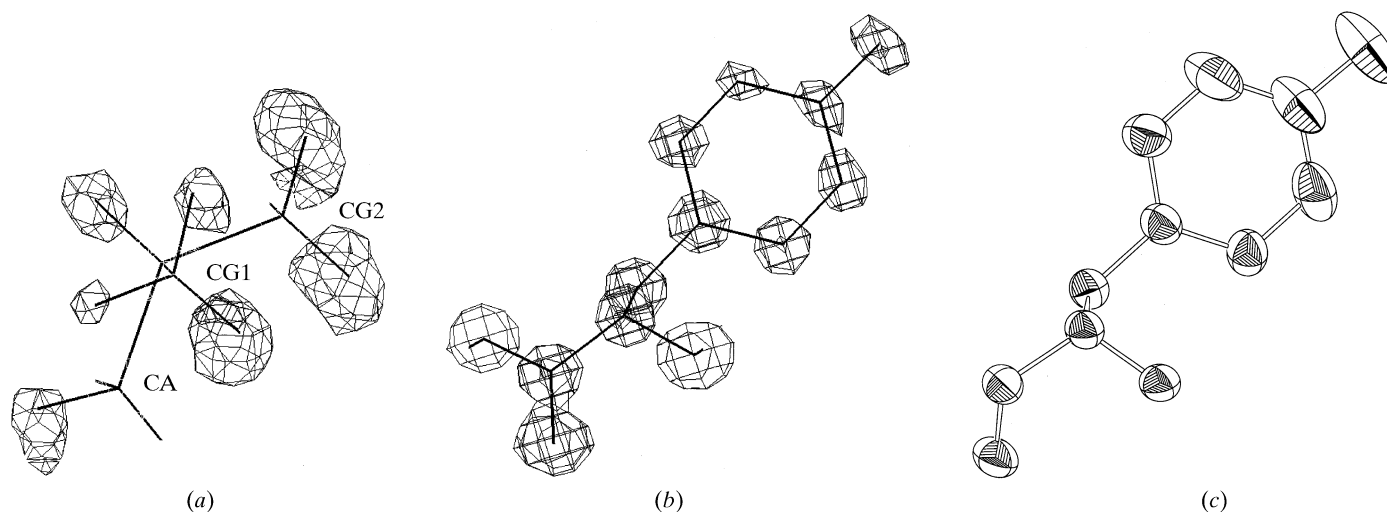


Fig. 18.4.5.1. (a), (b) Representative electron-density maps for the refinement of *Clostridium acidurici* ferredoxin at 0.94 Å resolution (Dauter, Wilson *et al.*, 1997). (a) The density for hydrogen atoms (at  $3\sigma$ ) omitted from the structure-factor calculation for Val42. (b) The  $(2F_o - F_c)$  density for Tyr30, contoured at  $3\sigma$ . (c) The thermal ellipsoids corresponding to (b), drawn at the 33% probability level using *ORTEPII* (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 *CCP4* suite (Collaborative Computational Project, Number 4, 1994). For rigid groups such as the NH amide, aromatic rings,  $-\text{CH}_2-$  or  $=\text{CH}-$ , the position is accurately defined by the bonding scheme. For groups such as methyl  $\text{CH}_3$  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*<sub>free</sub>. 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*<sub>free</sub> will not. Experience shows that at 2 Å or less there is no drop in *R*<sub>free</sub>, 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