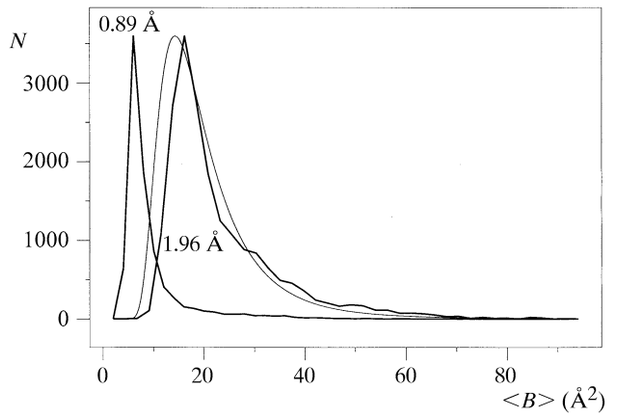

Figure 18.4.1.2
Histograms of B values for a protein structure, *Micrococcus lysodecticus* catalase (Murshudov *et al.*, 1999), for two different crystals which diffracted to different limiting resolutions. For both crystals, the resolution cutoff reflects the real diffraction limit from the sample,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

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

18.4.1.3. A theoretical approach to 'atomic resolution'

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

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

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

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

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

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

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

18.4.2. Data

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

18.4.2.1. Data quality

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

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

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

18.4.2.2. Anisotropic scaling

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

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

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

18.4.3. Computational algorithms and strategies

18.4.3.1. Classical least-squares refinement of small molecules

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

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

18.4.3.2. Least-squares refinement of large structures

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

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

A further simplification involves the conjugate-gradient method or the diagonal approximation to the normal matrix (the second derivative of minus the log of the likelihood function in the case of maximum likelihood), which essentially ignores all off-diagonal terms of the least-squares matrix. For the conjugate-