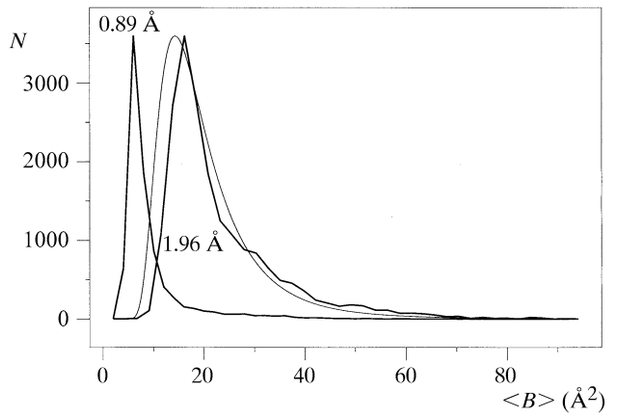

Figure 18.4.1.2
Histograms of B values for a protein structure, *Micrococcus lysodecticus* catalase (Murshudov *et al.*, 1999), for two different crystals which diffracted to different limiting resolutions. For both crystals, the resolution cutoff reflects the real diffraction limit from the sample,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