derived from databases (such as rotamer conformations; Jones *et al.*, 1991; Zou & Mowbray, 1994; Kleywegt & Jones, 1998) and careful rebuilding and refinement protocols (Kleywegt & Jones, 1997).

(6) Various types of error (possibly, to some extent, compensating ones) can be introduced during refinement, particularly if a satisfactorily low value for the conventional crystallographic R value is desired (Kleywegt & Jones, 1995*b*). This can always be achieved (even for models that have been deliberately traced backwards through the density; Jones *et al.*, 1991; Kleywegt & Jones, 1995*b*; Kleywegt & Brünger, 1996) by removing data that do not agree well with the model (through a resolution and σ cutoff), by not exploiting the redundancy of noncrystallographic symmetry properly (Kleywegt & Jones, 1995*b*; Kleywegt, 1996), by using an inappropriate temperature-factor model, by introducing alternative conformations and refining occupancies when these are not warranted by the information content of the data, by sprinkling the model with solvent molecules, and by reducing the weight given to the geometric and other restraints relative to the weight given to the crystallographic data.

One should realise that making errors is almost unavoidable (given the fact that one usually deals with limited resolution and less than perfect phases). The purpose of refinement and rebuilding is to detect and fix the errors to obtain the best possible final model that will be interpreted in terms of the biological role of the protein. Nevertheless, sometimes errors do persist into the publication and the deposited model. This may be a consequence of factors such as (Jones & Kjeldgaard, 1997):

(1) inexperienced, under-supervised people who do the work (and have a supervisor who may be in a hurry to publish);

(2) computer programs used as black boxes;

(3) new methods not adopted until the limitations of older ones have been experienced;

(4) intermediate models not subjected to critical and systematic quality analysis;

(5) use of 'quality indicators' that are strongly correlated with parameters that are restrained during refinement (r.m.s. deviation of bond lengths and angles from ideal values, r.m.s. ΔB for bonded atoms *etc.*).

21.1.3. Detecting outliers

21.1.3.1. *Classes of quality indicators*

Many statistics, methods and programs were developed in the 1990s to help identify errors in protein models. These methods generally fall into two classes: one in which only coordinates and B factors are considered (such methods often entail comparison of a model to information derived from structural databases) and another in which both the model and the crystallographic data are taken into account. Alternatively, one can distinguish between methods that essentially measure how well the refinement program has succeeded in imposing restraints (*e.g.* deviations from ideal geometry, conventional R value) and those that assess aspects of the model that are 'orthogonal' to the information used in refinement (*e.g.* free R value, patterns of non-bonded interactions, conformational torsion-angle distributions). An additional distinction can be made between methods that provide overall (global) statistics for a model (such methods are suitable for monitoring the progress of the refinement and rebuilding process) and those that provide information at the level of residues or atoms (such methods are more useful for detecting local problems in a model). It is important to realise that almost all coordinate-based validation methods detect *outliers* (*i.e.* atoms or residues with unusual properties): to assess whether an outlier arises from an *error* in the model or whether it is a genuine, but unusual, *feature* of the structure, one must inspect the (preferably unbiased) electron-density maps (Jones *et al.*, 1996)!

In this section, some quality indicators will be discussed that have been found to be particularly useful in daily protein crystallographic practice for the purpose of detecting problems in intermediate models. Section 21.1.7 provides a more extensive discussion of many of the quality criteria that are or have been used by macromolecular crystallographers.

21.1.3.2. *Local statistics*

From a practical point of view, these are the most useful for the crystallographer who is about to rebuild a model. Examples of useful quality indicators are:

(1) The real-space fit (Jones *et al.*, 1991; Chapman, 1995; Jones & Kjeldgaard, 1997; Vaguine *et al.*, 1999), expressed as an R value or as a correlation coefficient between 'observed' and calculated density. This property can be calculated for any subset of atoms, *e.g.* for an entire residue, for main-chain atoms or for side-chain atoms. It is best to use a map that is biased by the model as little as possible [*e.g.*, a σ_A-weighted map (Read, 1986), an NCS-averaged map (Kleywegt & Read, 1997) or an omit map (Bhat & Cohen, 1984; Hodel *et al.*, 1992)]. In practice, the real-space fit is strongly correlated with the atomic temperature factors, even though these are not used in the calculations.

(2) The Ramachandran plot (Ramakrishnan & Ramachandran, 1965; Kleywegt & Jones, 1996*b*). Residues with unusual main-chain φ, ψ torsion-angle combinations that do not have unequivocally clear electron density are almost always in error. However, one should keep in mind that the error may have its origin in (one of) the neighbouring residues. For instance, if the peptide O atom of a residue is pointing in the wrong direction, the φ value for the next residue may be off by 150–180° (Kleywegt, 1996; Kleywegt & Jones, 1998).

(3) The pep-flip value (Jones *et al.*, 1991; Kleywegt & Jones, 1998). This statistic measures the r.m.s. distance between the peptide O atom of a residue and its counterparts found in a database of well refined high-resolution structures that occur in parts of those structures with a similar local C^α backbone conformation. If the pep-flip value is large (*e.g.* >2.5 Å), the residue is termed an outlier, but whether it is an error can only be determined by inspecting the local density.

(4) The rotamer side-chain fit value (Jones *et al.*, 1991; Kleywegt & Jones, 1998). This statistic measures the r.m.s. distance between the side-chain atoms of a residue and those in the most similar rotamer conformation for that residue type. A value greater than ∼1.0–1.5 Å signals an outlier. In many cases (particularly, but not exclusively, at low resolution), a non-rotamer side chain can easily be replaced by a rotamer conformation, perhaps in conjunction with a slight rigid-body movement of the entire residue or with some adjustment of the side-chain torsion angles (Zou & Mowbray, 1994; Kleywegt & Jones, 1997).

(5) Hydrogen-bonding analysis. The correct orientation of histidine, asparagine and glutamine side chains cannot usually be inferred from electron density alone. Inexperienced crystallographers can benefit from suggestions based on the analysis of hydrogen-bonding networks (Hooft *et al.*, 1996*b*), although every case should be examined critically (*e.g.* the program does not know about solvent molecules that have not yet been added to the model or that cannot be placed because of the limitations of the data; in addition, sometimes an amino group may be interacting with an aromatic side chain).

In addition to these criteria, residues with other unusual features should be examined in the electron-density maps for the crystallographer to be able to decide whether they are in error. Such features may pertain to unusual temperature factors, unusual

occupancies, unusual bond lengths or angles, unusual torsion angles or deviations from planarity (*e.g.* for the peptide plane), unusual chirality (*e.g.* for the C^{α} atom of every residue type except glycine), unusual differences in the temperature factors of chemically bonded atoms, unusual packing environments (Vriend & Sander, 1993), very short distances between non-bonded atoms (including symmetry mates), large positional shifts during refinement, unusual deviations from noncrystallographic symmetry (Kleywegt & Jones, 1995*b*; Kleywegt, 1996) *etc.*

21.1.3.3. *Global statistics*

The crystallographic *R* value used to be the major global quality indicator until it was realised that it can easily be fooled, especially at low resolution (Brändén & Jones, 1990; Jones *et al.*, 1991; Brünger, 1992*a*; Kleywegt & Jones, 1995*b*). The free *R* value, introduced by Brünger (1992*a*, 1993), has been shown to be much more reliable and harder to manipulate (Kleywegt & Brünger, 1996; Brünger, 1997). It is excellently suited for monitoring the progress of refinement, for detecting major problems with model or data and for helping reduce over-fitting of the data (which occurs if many more parameters are refined in a model than is warranted by the information content of the crystallographic data). Moreover, the free *R* value can be used to estimate the coordinate error of the final model (Kleywegt *et al.*, 1994; Kleywegt & Brünger, 1996; Brünger, 1997; Cruickshank, 1999).

In addition, the average or r.m.s. values for many of the local statistics, their minimum or maximum values or the percentage of outliers can be quoted and used to obtain an impression of the overall quality of the model and the overall fit of the model to the data.

21.1.4. Fixing errors

The object of model rebuilding is generally twofold: (1) to make the model as complete and detailed as the data will allow one to do confidently (*e.g.* to add previously unmodelled loops, ligands, water molecules *etc.*) and (2) to remove errors. At first glance, it may not seem all that important to fix each and every side chain and to correct all peptide O atoms that are pointing in the wrong direction, but one should keep in mind that an error in the scattering factor (atom type or charge), position or *B* factor of even a single atom will be detrimental to the entire model. Particularly in the early stages of model rebuilding and refinement, one often finds that after an extensive round of rebuilding followed by more refinement, the density improves dramatically and new features become clear. One should also keep in mind that incorrect features of a model may be very persistent and become 'self-fulfilling prophecies', a phenomenon known as 'model bias' (Ramachandran & Srinivasan, 1961; Read, 1986, 1994, 1997; Hodel *et al.*, 1992). This is particularly relevant in cases where unbiased phase information (*e.g.* SIRAS, MIR or MAD phases, or phases obtained after NCS or multiple-crystal averaging) is not available.

For error detection to be effective, it is best not to approach the rebuilding process in a haphazard way (Kleywegt & Jones, 1997). *O* users can employ a program called *OOPS* (Kleywegt & Jones, 1996*a*) to carry out this task in a systematic yet convenient fashion. This program uses information calculated by *O* (*e.g.* pep-flip and real-space fit values) and retrieves or derives other information from a PDB file of the current model (*e.g.* temperature factors, Ramachandran plot, changes with respect to the previous model). Moreover, results from a coordinate-based quality check by the *WHAT IF* program (Vriend, 1990; Hooft *et al.*, 1996) can be included. In all, several dozen quality indicators can be used and plots and statistics for many of these can be produced by the program. The program's most useful feature, however, is that it will generate *O* macros that when executed in *O* will take the crystallographer on a journey to all the residues that may require attention because they are outliers for one or more quality criteria. This makes the rebuilding process often faster and certainly more efficient and focused than a residue-by-residue walk through the model. In addition, it teaches inexperienced crystallographers to recognize and diagnose common model errors.

If a residue is an outlier for a certain criterion, the crystallographer has to inspect the local density and the structural context and decide the course of action. If the residue is in a region of the model in which many residues are outliers for many criteria, there may be something seriously wrong locally (for instance, there could be a register error), possibly because the density is poor. If there is poor density for several residues in a row, the crystallographer might consider leaving these residues out of the model for the next refinement round or cutting off the side chains at the C^{β} atoms. Sometimes local errors are correlated, such as a pep-flip error in one residue and a Ramachandran violation for its C-terminal neighbour, or a residue with a non-rotamer conformation and high temperature factors in conjunction with a poor real-space fit. *O* contains many tools to manipulate individual residues and atoms (Jones *et al.*, 1991; Jones & Kjeldgaard, 1994, 1997; Kleywegt & Jones, 1997), *e.g.* to flip a peptide plane, to replace a side chain by a rotamer conformation, to change side-chain torsion angles in order to optimize the fit to the density, to move groups of atoms, to use real-space refinement on a single residue or a zone of residues, to 'mutate' a residue to alanine *etc*. Together they constitute a toolbox with which many problems, once recognized, can be fixed relatively effortlessly (Kleywegt & Jones, 1997).

21.1.5. Preventing errors

As with everything else, when it comes to building a model of a protein, prevention of errors is the best medicine. Some general guidelines can be given (Dodson *et al.*, 1996; Kleywegt & Jones, 1997).

(1) Try to obtain the best possible set of data and the best possible set of phases for those data. If the structure has noncrystallographic symmetry (or if multiple crystal forms are available), use electron-density averaging to remove model bias and to reduce phase errors (Kleywegt & Read, 1997). In the absence of noncrystallographic symmetry, use maps that are biased by the model as little as possible [*e.g.* σ_A-weighted (Read, 1986) or omit maps (Bhat & Cohen, 1984; Bhat, 1988; Hodel *et al.*, 1992)]. If experimental phase information is available, keep and consult the experimental map(s). Experimental phases can also be used throughout the refinement process to alleviate or prevent some problems.

(2) Use databases to construct the initial model (or new parts of the model; Jones *et al.*, 1991; Kleywegt & Jones, 1998). All the crystallographer needs to do is to roughly place the C^{α} atoms in the density. The model-building program can then 'recycle' well refined high-resolution structures to place the main-chain atoms. Similarly, side-chain conformations should initially be chosen from the set of preferred rotamers for each residue type, perhaps in combination with a rigid-body rotation of the entire residue around its C^{α} atom and/or with minor adjustment of the torsion angles of long side chains (arginine, lysine *etc*.).

(3) After every cycle of refinement, carry out a critical analysis of the quality of the current model. This entails the calculation of properties such as those discussed in Section 21.1.3 and the inspection of the residues that are outliers for any of them, as described in Section 21.1.4. Be conservative during rebuilding, especially when the model is incomplete and possibly full of errors.

(4) Design a refinement protocol that is appropriate for the available data. If NCS restraints do not give a significantly better

references