International Tables for Crystallography (2006). Vol. F, Chapter 21.1, pp. 497–506.

21. STRUCTURE VALIDATION

21.1. Validation of protein crystal structures

BY G. J. KLEYWEGT

21.1.1. Introduction

Owing to the limited resolution and imperfect phase information that macromolecular crystallographers usually have to deal with, building and refining a protein model based on crystallographic data is not an exact science. Rather, it is a subjective process, governed by experience, prejudices, expectations and local practices (Brändén & Jones, 1990; Kleywegt & Jones, 1995*b*, 1997). This means that errors in this process are almost unavoidable, but it is the crystallographer's task to remove as many of these as possible prior to analysis, publication and deposition of the structure. With high-resolution data and good phases, the resulting model is probably more than 95% a consequence of the data, although even at atomic resolution, subjective choices must still be made: which refinement program to use, whether to include alternative conformations, whether to model explicit H atoms, how to model temperature factors, which restraints and constraints to apply, which peaks in the maps to interpret as solvent molecules and how to treat noncrystallographic symmetry (NCS). Once the resolution becomes worse than ∼2 Å, this balance shifts and some published protein models appear to have been determined more by some crystallographer's imagination than by any experimental data.

Subjectivity is not necessarily a problem, provided that the crystallographer is experienced, knows what he or she is doing and is aware of the limitations that the experimental data impose on the model. However, even inexperienced people can avoid many of the pitfalls of model building and refinement. Supervisors have a major responsibility in this respect: education is an important factor (Dodson *et al.*, 1996). Students who have built and refined a previously determined structure from scratch as a training exercise will have met most of the problems that can be encountered in real life (Jones & Kjeldgaard, 1997). Apart from hands-on experience, there are many other methods to reduce or avoid errors. These include (1) the use of information derived from databases of well refined structures in model building (Kleywegt & Jones, 1998) [*e.g.* to generate main-chain coordinates from a C^{α} trace (Jones & Thirup, 1986) and side-chain coordinates from preferred rotamer conformations (Ponder & Richards, 1987)]; (2) the use of various sorts of local quality checks (to detect residues that for one or more reasons are deemed 'unusual' and that require further scrutiny and perhaps adjustment; Kleywegt & Jones, 1996*a*, 1997); and (3) the use of global quality indicators [*e.g.* the use of the free *R* value (Brünger, 1992*a*, 1993) to signal major errors, to prevent overfitting, and to monitor the progress of the rebuilding and refinement process (Kleywegt & Jones, 1995*b*; Kleywegt & Brünger, 1996; Brünger, 1997)].

21.1.2. Types of error

At every step of a crystal structure determination, the danger of making mistakes looms (Brändén & Jones, 1990; Janin, 1990). In this laboratory, for instance, a protein other than that intended was once purified (lysozyme instead of cellular retinoic acid binding protein), which obviously made the molecular-replacement problem rather intractable. Similarly, there is at least one published crystallization report of a protein other than for which the crystallographers had hoped: crystals of the light-harvesting complex LH1 actually turned out to be of bacterioferritin (Nunn *et al.*, 1995). There is at least one case in which an incorrect molecular-replacement solution was found that persisted all the way to the final published model, namely that of turkey egg-white lysozyme (Bott & Sarma, 1976). During data collection and processing, space-group assignment errors are occasionally made, such as in the case of chloromuconate cycloisomerase (Hoier *et al.*, 1994; Kleywegt *et al.*, 1996). A more common problem at this stage, however, is weak and/or incomplete data. The importance of complete data sets with a high signal-to-noise ratio and high redundancy for the success of the subsequent structure determination process (phasing, model building and refinement) cannot be overstressed. However, the discussion in this chapter will focus mainly on errors that may creep into a protein model built and refined by a crystallographer. Such errors come in various classes (Brändén & Jones, 1990) and, fortunately, the frequency of each type of error is inversely proportional to its seriousness.

(1) In the worst case, the model (or a sub-unit) may essentially be completely wrong. Recently identified examples of this type of problem include asparaginase/glutaminase (Ammon *et al.*, 1988; Lubkowski *et al.*, 1994) and photoactive yellow protein (McRee *et al.*, 1989; Borgstahl *et al.*, 1995).

(2) In other cases, secondary-structure elements may have been correctly identified for the most part, but incorrectly connected. This happened, for instance, in the structure determination of D-Ala-D-Ala carboxypeptidase/transpeptidase (Kelly *et al.*, 1986; Kelly & Kuzin, 1995).

(3) A fairly common mistake during the initial tracing is to overlook one residue, which leads to a register error (or frame shift). The model is usually brought back into register with the density a bit further down the sequence, where the opposite error is made (*e.g.* an extra residue is inserted into density for a turn). This is a serious error, but it is usually possible to detect and correct it in the course of the refinement and rebuilding process (Kleywegt *et al.*, 1997). However, it is not impossible for such an error to persist, particularly in low-resolution studies. Indeed, in one case in which a published 3.0 Å structure was re-refined, a register error was detected involving about two dozen residues (Hoier *et al.*, 1994; Kleywegt *et al.*, 1996).

(4) Sometimes the primary sequence used by the crystallographer contains one or more mistakes. These may arise from post-translational modifications, from sequencing errors, from the absence of a published amino-acid sequence at the time of tracing, from unanticipated cloning artifacts or simply from trivial 'transcription artifacts'. In this laboratory, the latter occurred during the refinement of human α-class glutathione S-transferase A1-1 (Sinning *et al.*, 1993), where one glycine residue had mistakenly been typed in as aspartate. Fortunately, the error revealed itself even at low resolution (2.6 Å), because the model was refined conservatively. In this case, the group of side-chain atoms obtained a very high *B* factor, in contrast to the very low *B* factor for the grouped main-chain atoms.

(5) The most common type of model-building error is locally incorrect main-chain and/or side-chain conformations. Such errors are easy to make in low-resolution maps calculated with imperfect phases. Moreover, multiple conformations are often unresolved even at moderately high resolution (∼2 Å), which further complicates the interpretation of side-chain density. Nevertheless, many of them can be avoided through the use of information

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

free R value than NCS constraints, then use constraints. If NCS restraints are to be employed, then use the experimental map to design a suitable NCS-restraint scheme (Kleywegt, 1999). Avoid the temptation to model alternative conformations in low-resolution maps or to place putative solvent molecules in every local maximum of the ($F_o - F_c$, α_c) difference map. In other words, be conservative and remember that the maxim 'where freedom is given, liberties are taken' is highly applicable to refinement programs (Hendrickson & Konnert, 1980; Kleywegt & Jones, 1995*b*).

(5) Adopt methodological advances as soon as they become available. Several innovations have only been slowly accepted by the mainstream (*e.g.* the use of databases in building and rebuilding, the use of the free R value, the use of electron-density averaging in molecular-replacement cases, bulk-solvent modelling). The most prominent recent development is the use of likelihood-based refinement programs (Bricogne & Irwin, 1996; Pannu & Read, 1996; Murshudov *et al.*, 1997; Adams *et al.*, 1997; Pannu *et al.*, 1998). These programs produce better models and maps and considerably reduce over-fitting (as assessed by the difference between the free and conventional R values).

(6) Most importantly, the crystallographer should be hyper-critical towards the fruits of his or her own labour. Every intermediate model is a hypothesis to be shot down (Jones & Kjeldgaard, 1994). The crystallographer should be more critical than the supervisor, the supervisor more critical than the referee and the referee more critical than the casual reader. It goes without saying that the reader, casual or not, should have access to model coordinates, experimental data and electron-density maps!

21.1.6. Final model

Once the refinement is finished [*i.e.* once the ($F_o - F_c$, α_c) difference map is featureless (Cruickshank, 1950) and parameter shifts in further refinement cycles are negligibly small], three tasks remain: validation of the final model, description and analysis of the structure, and deposition of the model coordinates and the crystallographic data with the Protein Data Bank (Bernstein *et al.*, 1977).

Until a few years ago, validation of the final model typically entailed calculating the conventional R value, r.m.s. deviations from ideal values of bond lengths and angles, average temperature factors, and a Luzzati-type estimate of coordinate error. Kleywegt & Jones (1995*b*) showed that these statistics are not necessarily even remotely related to the actual quality of a model. Based on these criteria, a backwards-traced protein model was of higher apparent quality than a carefully refined correct model. After this, the realisation sunk in that the best validation criteria are those that assess aspects of the model that are 'orthogonal' to the information used during model refinement and rebuilding. For instance, the main-chain φ and ψ torsion angles are usually not restrained during refinement; this makes the Ramachandran plot such a powerful validation tool (Kleywegt & Jones, 1996*b*, 1998). Other examples of useful independent tests include the profile method of Eisenberg and co-workers (Lüthy *et al.*, 1992), the directional atomic contact analysis method of Vriend & Sander (1993) and the threading-potential method of Sippl (1993).

In general, all quality checks provide necessary, but in themselves insufficient, indications as to whether or not a model is essentially correct. A truly good model should make sense with respect to what is currently known about physics, chemistry, crystallography, protein structures, statistics and (last, but not least) biology and biochemistry (Kleywegt & Jones, 1995*a*). A good model will typically score well on most if not all validation criteria, whereas a poor one will score poorly on many criteria. The same is true at the level of residues: a poor or erroneous region in a model will be characterized by violations of many residue-level quality criteria (Kleywegt & Jones, 1997).

21.1.7. A compendium of quality criteria

In this section, some of the quality and validation criteria that have been used by macromolecular crystallographers are summarized (for more detailed information, the reader is referred to the primary literature). When judging how useful or powerful these criteria are in a certain case, one should keep in mind that any criterion that has been used explicitly or implicitly during model refinement (*e.g.* geometric restraints) or rebuilding (*e.g.* rotamer libraries) does *not* provide a truly independent check on the quality of the model.

Many, but not all, of the criteria discussed below pertain specifically to protein models. Comparatively little work has been performed on the validation of nucleic acid models, although there are indications that there is a need for such procedures (Schultze & Feigon, 1997). The situation would appear to be even worse for hetero-entities (*e.g.* ligands, inhibitors, cofactors, covalent attachments, saccharides, metals, ions; van Aalten *et al.*, 1996; Kleywegt & Jones, 1998).

21.1.7.1. *Data quality*

Although many quality and validation criteria have been developed for assessing coordinate sets of protein models, comparatively few criteria are available for assessing the quality of the crystallographic data.

21.1.7.1.1. *Merging R values*

Possibly the most common mistake in papers describing protein crystal structures is an incorrectly quoted formula for the merging R value (calculated during data reduction),

$$R_{\rm merge} = \sum_h \sum_i |I_{h,\,i} - \langle I_h \rangle| / \sum_h \sum_i I_{h,\,i},$$

where the outer sum (h) is over the unique reflections (in most implementations, only those reflections that have been measured more than once are included in the summations) and the inner sum (i) is over the set of independent observations of each unique reflection (Drenth, 1994). This statistic is supposed to reflect the spread of multiple observations of the intensity of the unique reflections (where the multiple observations may derive from symmetry-related reflections, different images or different crystals). Unfortunately, $R_{\rm merge}$ is a very poor statistic, since its value increases with increasing redundancy (Weiss & Hilgenfeld, 1997; Diederichs & Karplus, 1997), even though the signal-to-noise ratio of the average intensities will be higher as more observations are included (in theory, an N-fold increase of the number of independent observations should improve the signal-to-noise ratio by a factor of $N^{1/2}$). At high redundancy, the value of $R_{\rm merge}$ is directly related to the average signal-to-noise ratio (Weiss & Hilgenfeld, 1997): $R_{\rm merge} \simeq 0.8/\langle I/\sigma(I) \rangle$.

Diederichs & Karplus (1997) have suggested a number of alternative measures that lack most of the drawbacks of $R_{\rm merge}$. Their statistic $R_{\rm meas}$ is similar to $R_{\rm merge}$, but includes a correction for redundancy (m),

$$R_{\rm meas} = \sum_h [m/(m-1)]^{1/2} \sum_i |I_{h,\,i} - \langle I_h \rangle| / \sum_h \sum_i I_{h,\,i}.$$

Another statistic, the pooled coefficient of variation (PCV), is defined as

$$\mathrm{PCV} = \sum_h \{[1/(m-1)] \sum_i (I_{h,\,i} - \langle I_h \rangle)^2\}^{1/2} / \sum_h \langle I_h \rangle.$$

Since PCV = $1/\langle I/\sigma(I)\rangle$, this quantity also provides an indication as to whether the standard deviations $\sigma(I)$ have been estimated appropriately. Finally, the statistic $R_{\rm mrgd\text{-}F}$, used for assessing the quality of the reduced data, enables a direct comparison of this merging R value with the refinement residuals R and $R_{\rm free}$.

Ideally, merging statistics should be quoted for all resolution shells (which should not be too broad), as well as for the entire data set. However, as a minimum, the values for the two extreme (low- and high-resolution) shells and for the entire data set should be reported.

21.1.7.1.2. *Completeness*

Data completeness can be assessed by calculating what fraction of the unique reflections within a range of Bragg spacings that could in theory be observed has actually been measured. Ideally, completeness should be quoted for all resolution shells (which should not be too broad), as well as for the entire data set. However, as a minimum, the values for the two extreme (low- and high-resolution) shells and for the entire data set should be reported.

21.1.7.1.3. *Redundancy*

Redundancy is defined as the number of independent observations (after merging of partial reflections) per unique reflection in the final merged and symmetry-reduced data set. Ideally, average redundancy should be quoted for all resolution shells (which should not be too broad), as well as for the entire data set. However, as a minimum, the values for the two extreme (low- and high-resolution) shells and for the entire data set should be reported.

21.1.7.1.4. *Signal strength*

The average strength or significance of the observed intensities can be expressed in different ways. Values that are often quoted include the percentage of reflections for which $I/\sigma(I)$ exceeds a certain value (usually 3.0) and the average value of $I/\sigma(I)$. Ideally, these numbers should be quoted for all resolution shells (which should not be too broad), as well as for the entire data set. However, as a minimum, the values for the two extreme (low- and high-resolution) shells and for the entire data set should be reported.

21.1.7.1.5. *Resolution*

The nominal resolution limits of a data set are chosen by the crystallographer, usually at the data-processing stage, and ought to reflect the range of Bragg spacings for which useful intensity data have been collected. Unfortunately, owing to the subjective nature of this process, resolution limits cannot be compared meaningfully between data sets processed by different crystallographers. Careful crystallographers will take factors such as shell completeness, redundancy and $\langle I/\sigma(I)\rangle$ into account, whereas others may simply look up the minimum and maximum Bragg spacing of all observed reflections. Bart Hazes (personal communication) has suggested defining the effective resolution of a data set as that resolution at which the number of observed reflections would constitute a 100% complete data set. Alternatively, Vaguine *et al.* (1999) define the effective (or optical) resolution as the expected minimum distance between two resolved peaks in the electron-density map and calculate this quantity as $2\Delta_P/2^{1/2}$, where Δ_P is the width of the origin Patterson peak. One day, hopefully, the term 'resolution' will be replaced by an estimate of the information content of data sets. Randy Read (personal communication) has carried out preliminary work along these lines.

21.1.7.1.6. *Unit-cell parameters*

The accuracy of unit-cell parameters has been shown to be grossly overestimated for small-molecule crystal structures (Taylor & Kennard, 1986). Not intimidated by this observation, some macromolecular crystallographers routinely quote unit-cell axes of 100–200 Å with a precision of 0.01 Å. An analysis of several high-resolution protein crystal structures has revealed that surprisingly large errors in the unit-cell parameters appear to be quite common (at least if synchrotron sources are used for data collection; EU 3-D Validation Network, 1998). Such errors can be detected *a posteriori* by checking if the bond lengths in a model show any systematic, perhaps direction-dependent, variations from their target values.

21.1.7.1.7. *Symmetry*

From the symmetry of the diffraction pattern, the point-group symmetry of the crystal lattice can usually be derived. It is important to merge the data in the point group with the highest possible symmetry (usually assessed using merging statistics) in order to minimize the chance of making an incorrect space-group assignment (Marsh, 1995, 1997; Kleywegt *et al.*, 1996). Once the first data set has been processed, it is always useful to compute a self-rotation function. A non-origin peak of comparable strength to the origin peak will indicate that the true space group has higher symmetry. [Similarly, a self-Patterson function can be calculated at this stage to detect any purely translational NCS (Kleywegt & Read, 1997).] Once the final model is available, a search for possibly missed higher symmetry can be carried out, *e.g.* using the method developed by Hooft *et al.* (1994).

Sometimes crystallographic symmetry breaks down (pseudo-symmetry): an apparent higher symmetry at low resolution does not hold at higher resolution. In some cases, this is a consequence of the chemistry of the system studied (*e.g.* an asymmetric ligand bound by a symmetric protein dimer). In other cases, it may go undetected and complicate space-group determination and solution and refinement of the structure.

When it comes to space-group determination, many of the lessons learned by small-molecule crystallographers also apply to macromolecular crystallography (Marsh, 1995; Watkin, 1996).

21.1.7.2. *Model quality, coordinates*

Many criteria (and computer programs) are available to check for structural outliers based only on analysis of Cartesian coordinate sets.

21.1.7.2.1. *Geometry and stereochemistry*

The covalent geometry of a model can be assessed by comparing bond lengths and angles to a library of 'ideal' values. In the past, every refinement and modelling program had its own set of 'ideal' values. This even made it possible to detect (with 95% accuracy) with which program a model had been refined, simply by inspecting its covalent geometry (Laskowski, Moss & Thornton, 1993). Nowadays, standard sets of ideal bond lengths and bond angles derived from an analysis of small-molecule crystal structures from the CSD (Allen *et al.*, 1979) are available for proteins (Engh & Huber, 1991; Priestle, 1994) and nucleic acids (Parkinson *et al.*, 1996). For other entities, typical bond lengths and bond angles can be taken from tables of standard values (Allen *et al.*, 1987) or derived by other means (Kleywegt & Jones, 1998; Greaves *et al.*, 1999).

For bond lengths, the r.m.s. deviation from ideal values is invariably quoted. Deviations from ideality of bond angles can be expressed directly as an angular r.m.s. deviation or in terms of angle distances (*i.e.* the angle $\angle ABC$ is measured by the 1–3 distance $|AC|$; note that this distance is also implicitly dependent on the bond

lengths $|AB|$ and $|BC|$). There are some indications that protein geometry cannot always be captured by assuming unimodal distributions (*i.e.* geometric features with only a single 'ideal' value). For example, Karplus (1996) found that the main-chain bond angle τ_3 ($\angle$N—C^α—C) varies as a function of the main-chain torsion angles φ and ψ.

Chirality is another important criterion in the case of biomacromolecules: most amino-acid residues will have the L configuration for their C^α atom. Also, the C^β atoms of threonine and isoleucine residues are chiral centres (IUPAC–IUB Commission on Biochemical Nomenclature, 1970; Morris *et al.*, 1992). Chirality can be assessed in terms of improper torsion angles or chiral volumes. For example, to check if the C^α atom of any residue other than glycine has the L configuration, the improper (or virtual) torsion angle C^α—N—C—C^β should have a value of about +34° (a value near −34° would indicate a D-amino acid). The torsion angle is called improper or virtual because it measures a torsion around something other than a covalent bond, in this case the N—C 'virtual bond'. The chiral volume is defined as the triple scalar product of the vectors from a central atom to three attached atoms (Hendrickson, 1985). For instance, the chiral volume of a C^α atom is defined as

$$V_{C^\alpha} = (\mathbf{r}_N - \mathbf{r}_{C^\alpha}) \cdot [(\mathbf{r}_C - \mathbf{r}_{C^\alpha}) \times (\mathbf{r}_{C^\beta} - \mathbf{r}_{C^\alpha})],$$

where $\mathbf{r}_X$ is the position vector of atom X. It should be noted that the chiral volume also implicitly depends on the bond lengths and angles involving the four atoms.

Another issue to consider is that of moieties that are necessarily planar (*e.g.* carboxylate groups, phenyl rings; Hooft *et al.*, 1996*a*). Again, planarity can be assessed in two different ways: by inspecting a set of (possibly improper) torsion angles and calculating their r.m.s. deviation from ideal values (*e.g.* all ring torsions in a perfectly flat phenyl ring should be 0°) or by fitting a least-squares plane through each set of atoms and calculating the r.m.s. distance of the atoms to that plane. Note that for double bonds, *cis* and *trans* configurations cannot be distinguished by deviations from a least-squares plane, but they can be distinguished by an appropriately defined torsion angle.

21.1.7.2.2. *Torsion angles (dihedrals)*

The conformation of the backbone of every non-terminal amino-acid residue is determined by three torsion angles, traditionally called φ (C_{i-1}—N_i—C_i^α—C_i), ψ(N_i—C_i^α—C_i—N_{i+1}) and ω (C_i^α—C_i—N_{i+1}—C_{i+1}^α). Owing to the peptide bond's partial double-bond character, the ω angle is restrained to values near 0° (*cis*-peptide) and 180° (*trans*-peptide). *Cis*-peptides are relatively rare and usually (but not always) occur if the next residue is a proline (Ramachandran & Mitra, 1976; Stewart *et al.*, 1990). The average ω-value for *trans*-peptides is slightly less than 180° (MacArthur & Thornton, 1996), but surprisingly large deviations have been observed in atomic resolution structures (Sevcik *et al.*, 1996; Merritt *et al.*, 1998). The ω angle therefore offers little in the way of validation checks, although values in the range of ±20 to ±160° should be treated with caution in anything but very high resolution models. The φ and ψ torsion angles, on the other hand, are much less restricted, but it has been known for a long time that owing to steric hindrance there are several clearly preferred combinations of φ, ψ values (Ramakrishnan & Ramachandran, 1965). This is true even for proline and glycine residues, although their distributions are atypical (Morris *et al.*, 1992). Also, an overwhelming majority of residues that are not in regular secondary-structure elements are found to have favourable φ, ψ torsion-angle combinations (Swindells *et al.*, 1995). For these reasons, the Ramachandran plot (essentially a φ, ψ scatter plot) is an extremely useful indicator of model quality (Weaver *et al.*, 1990; Laskowski, MacArthur *et al.*, 1993; MacArthur & Thornton, 1996; Kleywegt & Jones, 1996*b*; Kleywegt, 1996; Hooft *et al.*, 1997). Residues that have unusual φ, ψ torsion-angle combinations should be scrutinized by the crystallographer. If they have convincing electron density, there is probably a good structural or functional reason for the protein to tolerate the energetic strain that is associated with the unusual conformation (Herzberg & Moult, 1991). As a rule, the residue types that are most often found as outliers are serine, threonine, asparagine, aspartic acid and histidine (Gunasekaran *et al.*, 1996; Karplus, 1996). The quality of a model's Ramachandran plot is most convincingly illustrated by a figure. Alternatively, the fraction of residues in certain predefined areas of the plot (*e.g.* core regions) can be quoted, but in that case it is important to indicate which definition of such areas was used. Sometimes, one may also encounter a Balasubramanian plot, which is a linear φ, ψ plot as a function of the residue number (Balasubramanian, 1977).

In protein structures, the plane of the peptide bond can have two different orientations (approximately related by a 180° rotation around the virtual C^α—C^α bond) that are both compatible with a *trans* configuration of the peptide (Jones *et al.*, 1991). The correct orientation can usually be deduced from the density of the carbonyl O atom or from the geometric requirements of regular secondary-structure elements (in α-helices, all carbonyl O atoms point towards the C-terminus of the helix; in β-strands, carbonyl O atoms usually alternate their direction). In other cases, *e.g.* in loops with poor density, the correct orientation may be more difficult to determine and errors are easily made. By comparing the local C^α conformation to a database of well refined high-resolution structures, unusual peptide orientations can be identified and, if required, corrected (through a 'peptide flip'; Jones *et al.*, 1991; Kleywegt & Jones, 1997, 1998). Since flipping the peptide plane between residues i and $i+1$ changes the ψ angle of residue i and the φ angle of residue $i+1$ by ∼180°, erroneous peptide orientations may also lead to outliers in the Ramachandran plot (Kleywegt, 1996; Kleywegt & Jones, 1998).

All amino-acid residues whose side chain extends beyond the C^β atom contain one or more conformational side-chain torsion angles, termed χ_1 (N—C^α—C^β—X^γ, where X may be carbon, sulfur or oxygen, depending on the residue type; if there are two γ atoms, the χ_1 torsion is calculated with reference to the atom with the lowest numerical identifier, *e.g.* $O^{\gamma 1}$ for threonine residues), χ_2 (C^α—C^β—X^γ—X^δ) *etc.* Early on, it was found that the values that these torsion angles assume in proteins are similar to those expected on the basis of simple energy calculations and that in addition certain combinations of χ_1, χ_2 values are clearly preferred (so-called rotamer conformations; Janin *et al.*, 1978; James & Sielecki, 1983; Ponder & Richards, 1987). Analogous to Ramachandran plots, χ_1, χ_2 scatter plots can be produced that show how well a protein's side-chain conformations conform to known preferences (Laskowski, MacArthur *et al.*, 1993; Carson *et al.*, 1994). Alternatively, a score can be computed for each residue that shows how similar its side-chain conformation is to that of the most similar rotamer for that residue type. This score can be calculated as an r.m.s. distance between corresponding side-chain atoms (Jones *et al.*, 1991; Zou & Mowbray, 1994; Kleywegt & Jones, 1998) or it can be expressed as an r.m.s. deviation of side-chain torsion-angle values from those of the most similar rotamer (Noble *et al.*, 1993).

Other torsion angles that have been used for validation purposes include the proline φ torsion (restricted to values near −65° owing to the geometry of the pyrrolidine ring; Morris *et al.*, 1992) and the χ_3 torsion in disulfide bridges (defined by the atoms C^β—S—S′—$C^{\beta\prime}$ and restricted to values near +95 and −85°; Morris *et al.*, 1992). In addition to the torsion-angle values of individual residues, pooled standard deviations of χ_1 and/or χ_2 torsions have been used for validation purposes (Morris *et al.*, 1992; Laskowski, MacArthur *et al.*, 1993).

To assess the 'geometric strain' in a model on a per-residue basis, the refinement program *X-PLOR* (Brünger, 1992*b*) can produce geometric pseudo-energy plots. In such a plot, the ratio of $E_{\rm geom}(i)$/r.m.s.($E_{\rm geom}$) is calculated as a function of the residue number i. The pseudo-energy term $E_{\rm geom}$ consists of the sums of the geometric and stereochemical pseudo-energy terms of the force field ($E_{\rm geom} = E_{\rm bonds} + E_{\rm angles} + E_{\rm dihedrals} + E_{\rm impropers}$), involving only the atoms of each residue.

It has been observed that the more high-resolution protein structures become available, the more 'well behaved' proteins turn out to be, *i.e.* the distributions of conformational torsion angles and torsion-angle combinations become even tighter than observed previously and the numerical averages tend to shift somewhat (Ponder & Richards, 1987; Kleywegt & Jones, 1998; EU 3-D Validation Network, 1998; MacArthur & Thornton, 1999; Walther & Cohen, 1999).

21.1.7.2.3. *C^α-only models*

Validation of C^α-only models may be necessary if such a model is retrieved from the PDB to be used in molecular replacement or homology modelling exercises; however, not many validation tools can handle such models (Kleywegt, 1997). The C^α backbone can be characterized by C^α—C^α distances (∼2.9 Å for a *cis*-peptide and ∼3.8 Å for a *trans*-peptide), C^α—C^α—C^α pseudo-angles and C^α—C^α—C^α—C^α pseudo-torsion angles (Kleywegt, 1997). The pseudo-angles and torsion angles turn out to assume certain preferred value combinations (Oldfield & Hubbard, 1994), much like the backbone φ and ψ torsions, and this can be employed for the validation of C^α-only models (Kleywegt, 1997). In addition to these straightforward methods, the mean-field approach of Sippl (1993) is also applicable to C^α-only models.

21.1.7.2.4. *Contacts and environments*

Hydrophobic, electrostatic and hydrogen-bonding interactions are the main stabilizing forces of protein structure. This leads to packing arrangements where hydrophobic residues tend to interact with each other, where charged residues tend to be involved in salt links and where hydrophilic residues prefer to interact with each other or to point out into the bulk solvent. Serious model errors will often lead to violations of such simple rules of thumb and introduce non-physical interactions (*e.g.* a charged arginine residue located inside a hydrophobic pocket; Kleywegt *et al.*, 1996) that serve as good indicators of model errors. Directional atomic contact analysis (Vriend & Sander, 1993) is a method in which these empirical notions have been formalized through database analysis. For every group of atoms in a protein, it yields a score which in essence expresses how 'comfortable' that group is in its environment in the model under scrutiny (compared with the expectations derived from the database). If a region in a model (or the entire model) has consistently low scores, this is a very strong indication of model errors. The *ERRAT* program is based on the same principle, but it is less specific in that it assesses only six types of non-bonded interactions (CC, CN, CO, NN, NO and OO; Colovos & Yeates, 1993).

Hydrogen-bonding analysis can often be used to determine the correct orientation of asparagine, glutamine and histidine residues (McDonald & Thornton, 1995). Similarly, an investigation of unsatisfied hydrogen-bonding potential can be used for validation purposes (Hooft *et al.*, 1996*b*), as can calculation of hydrogen-bonding energies (Morris *et al.*, 1992; Laskowski, MacArthur *et al.*, 1993).

Finally, a model should not contain unusually short non-bonded contacts. Although most refinement programs will restrain atoms from approaching one another too closely, if any serious violations remain they are worth investigating, since they may signal an underlying problem (*e.g.* erroneous omission of a disulfide restraint or incorrect side-chain assignment).

21.1.7.2.5. *Noncrystallographic symmetry*

Molecules that are related by noncrystallographic symmetry exist in very similar, but not identical, physical environments. This implies that their structures are expected to be quite similar, although different relative domain orientations and local variations may occur (*e.g.* owing to different crystal-packing interactions; Kleywegt, 1996). Many criteria have been developed to quantify the differences between (NCS) related models. Some, such as the r.m.s. distance (*e.g.* on all atoms, backbone atoms or C^α atoms) are based on distances between equivalent atoms, measured after a (to some extent arbitrary; Kleywegt, 1996) structural superpositioning operation has been performed. Others are based on a comparison of torsion angles, be it of main-chain φ, ψ angles [*e.g.* $\Delta\varphi$, $\Delta\psi$ plot (Korn & Rose, 1994); multiple-model Ramachandran plot (Kleywegt, 1996); $\sigma(\varphi)$, $\sigma(\psi)$ plot (Kleywegt, 1996); circular variance (Allen & Johnson, 1991) plots of φ and ψ (G. J. Kleywegt, unpublished results); Euclidian φ, ψ distances (Carson *et al.*, 1994) or pseudo-energy values (Carson *et al.*, 1994)] or side-chain χ_1, χ_2 angles [*e.g.* multiple-model χ_1, χ_2 plot (Kleywegt, 1996); $\sigma(\chi_1)$, $\sigma(\chi_2)$ plots (Kleywegt, 1996); circular variance (Allen & Johnson, 1991) plots of χ_1 and χ_2 (G. J. Kleywegt, unpublished results); Euclidian χ_1, χ_2 distances (Carson *et al.*, 1994) or pseudo-energy values (Carson *et al.*, 1994)]. Still other methods are based on analysing differences in contact-surface areas (Abagyan & Totrov, 1997), temperature factors (Kleywegt, 1996) or the geometry of the C^α backbone alone (Flocco & Mowbray, 1995; Kleywegt, 1996). Many of these methods can also be used to compare the structures of related molecules in different crystals or crystal forms (*e.g.* complexes, mutants).

21.1.7.2.6. *Solvent molecules*

Solvent molecules provide an excellent means of 'absorbing' problems in both the experimental data and the atomic model. Neither their position nor their temperature factor are usually restrained (other than by the data and restraints that prevent close contacts) and sometimes even their occupancy is refined. At a resolution of ∼2 Å, crystallographers tend to model roughly one water molecule for every amino-acid residue and at 1.0 Å resolution this number increases to ∼1.6 (Carugo & Bordo, 1999). When waters are placed, it should be ascertained that they can actually form hydrogen bonds, be it to protein atoms or to other water molecules. Considering that several ions that are isoelectronic with water (Na^+, NH_4^+) are often used in crystallization solutions, one should keep in mind the possibility that some entities that have been modelled as water molecules could be something else (Kleywegt & Jones, 1997). A method to check if water molecules could actually be sodium ions, based on the surrounding atoms, has been published (Nayal & Di Cera, 1996).

21.1.7.2.7. *Miscellaneous*

Many other coordinate-based methods for assessing the validity or correctness of protein models have been developed. These include the profile method of Eisenberg and co-workers (Bowie *et al.*, 1991; Lüthy *et al.*, 1992), the inspection of atomic volumes (Pontius *et al.*, 1996), and the use of threading and other potentials (Sippl, 1993; Melo & Feytmans, 1998; Maiorov & Abagyan, 1998). Some of these methods are described in more detail elsewhere in this volume. The program *WHAT IF* (Vriend, 1990) contains a large array of quality checks, many of which are not available in other programs, that span the spectrum from administrative checks to global quality indicators (Hooft *et al.*, 1996). During the refinement

process, coordinate shifts can be used as a rough indication of 'quality' or, rather, convergence (Carson *et al.*, 1994; Kleywegt & Jones, 1996*a*). Crude models tend to undergo much larger changes during refinement than models that are essentially correct and complete. Also at the residue level, large coordinate shifts indicate residues that are worth a closer look.

Laskowski *et al.* (1994) have formulated single-number geometrical quality criteria, which they dubbed '*G* factors' in analogy to crystallographic *R* values. These *G* factors combine the results of a number of quality checks (covalent geometry, main-chain and side-chain torsion angles *etc.*) in a single number.

21.1.7.3. *Model quality, temperature factors*

In crystallographic refinement, atomic displacement parameters (ADPs; often referred to as temperature factors or *B* factors) model the effects of static and dynamic disorder. Except at high resolution (typically better than 1.5 Å), where there are sufficient observations to warrant refinement of anisotropic temperature factors, ADPs are usually constrained to be isotropic. The isotropic temperature factor *B* of an atom is related to the atom's mean-square displacement $\langle \Delta r^2 \rangle$ according to $B = 8\pi^2 \langle \Delta r^2 \rangle / 3$. Compared with the atomic coordinates, there are usually comparatively few restraints on temperature factors during refinement. Therefore, particularly at low resolution, temperature factors often function as 'error sinks' (Read, 1990). They absorb not only the effects of static and dynamic disorder, but also of various kinds of model errors.

Compared with the wealth of statistics that can be used to check and validate coordinates, there are relatively few methods available to assess how reasonable a model's temperature factors are. One obvious check is to see how well the average temperature factor of the model matches the value calculated from the data, using either a Wilson plot (Wilson, 1949) or the Patterson origin peak (Vaguine *et al.*, 1999). Since the average temperature factor of a model is usually not restrained, this is a useful check that has been used on several occasions to justify high average *B* factors. One should keep in mind that a low average *B* factor, *per se*, is not necessarily an indication of high model quality. For instance, a backwards-traced protein model can have a considerably lower average *B* factor than a correct model at a similar resolution (Kleywegt & Jones, 1995*b*). Average (and minimum and maximum) temperature-factor values can also be listed separately for various groups of atoms (*e.g.* individual protein or nucleic acid molecules, ligands, solvent molecules). A simple plot of residue-averaged temperature factors as a function of residue number may reveal regions of the molecule that have consistently high *B* factors, which may be a consequence of problems in the model (Kleywegt *et al.*, 1996).

Other statistics pertain to the r.m.s. differences in *B* factors between atoms that are somehow related, for example through a chemical bond (r.m.s. $\Delta B_{\rm bonded}$), through a 1–3 interaction or through noncrystallographic symmetry (possibly after correcting for any differences between the average *B* factors of the NCS-related molecules). Sometimes these statistics are calculated separately for main-chain and side-chain atoms. If the *B* factors of such related atoms have been restrained to be similar during refinement, these checks do not provide a convincing indication of the quality of the model. On the other hand, the *B* factors of atoms that have non-bonded interactions are usually not restrained to be similar, which renders the r.m.s. *B*-factor difference between such atoms (r.m.s. $\Delta B_{\rm non\text{-}bonded}$) slightly more informative.

Since proteins tend to consist of a tightly packed core with more flexible regions at the surface, a radial *B*-factor plot (*i.e.* a plot of the average *B* factor of all atoms in a certain distance range from the centre of the molecule as a function of the distance) is expected to be shaped roughly like a half-parabola. Kuriyan & Weis (1991) used a ten-parameter isotropic rigid-molecule model of the mean-square atomic displacement (Schomaker & Trueblood, 1968). After obtaining values for the ten parameters (either by refinement against the structure-factor data or by fitting to the refined *B* factors of the model), the *B* factor of any atom can be calculated and depends only on its coordinates. They found that regions with large discrepancies between the refined and fitted *B* factors tend to be associated with errors or problems in a model.

Validation of anisotropic ADPs (Merritt, 1999), non-unit occupancies and H atoms, all of which are usually associated with high-resolution data, is still in its infancy. The validity of modelling anisotropic ADPs can be assessed by comparing the reduction of the conventional and free *R* values. If occupancies are used for multiple conformations of, for example, a side chain, the sum of the occupancies should be unity.

21.1.7.4. *Model versus experimental data*

21.1.7.4.1. *R values*

The traditional statistic used to assess how well a model fits the experimental data is the crystallographic *R* value,

$$R = \sum w\big||F_o| - k|F_c|\big| \Big/ \sum |F_o|.$$

This statistic is closely related to the standard least-squares crystallographic residual $\sum w(|F_o| - k|F_c|)^2$ and its value can be reduced essentially arbitrarily by increasing the number of parameters used to describe the model (*e.g.* by refining anisotropic ADPs and occupancies for all atoms) or, conversely, by reducing the number of experimental observations (*e.g.* through resolution and σ cutoffs) or the number of restraints imposed on the model. Therefore, the conventional *R* value is only meaningful if the number of experimental observations and restraints greatly exceeds the number of model parameters. In 1992, Brünger introduced the free *R* value ($R_{\rm free}$; Brünger, 1992*a*, 1993, 1997; Kleywegt & Brünger, 1996), whose definition is identical to that of the conventional *R* value, except that the free *R* value is calculated for a small subset of reflections that are not used in the refinement of the model. The free *R* value, therefore, measures how well the model predicts experimental observations that are not used to fit the model (cross-validation). Until a few years ago, a conventional *R* value below 0.25 was generally considered to be a sign that a model was essentially correct (Brändén & Jones, 1990). While this is probably true at high resolution, it was subsequently shown for several intentionally mistraced models that these can be refined to deceptively low conventional *R* values (Jones *et al.*, 1991; Kleywegt & Jones, 1995*b*; Kleywegt & Brünger, 1996). Brünger suggests a threshold value of 0.40 for the free *R* value, *i.e.* models with free *R* values greater than 0.40 should be treated with caution (Brünger, 1997). Tickle and coworkers have developed methods to estimate the expected value of $R_{\rm free}$ in least-squares refinement (Tickle *et al.*, 1998). Since the difference between the conventional and free *R* value is partly a measure of the extent to which the model over-fits the data (*i.e.* some aspects of the model improve the conventional but not the free *R* value and are therefore likely to fit noise rather than signal in the data), this difference $R_{\rm free} - R$ should be small (Kleywegt & Jones, 1995*a*; Kleywegt & Brünger, 1996). Alternatively, the $R_{\rm free}$ ratio (defined as $R_{\rm free}/R$; Tickle *et al.*, 1998) should be close to unity. Various practical aspects of the use of the free *R* value have been discussed by Kleywegt & Brünger (1996) and by Brünger (1997).

Self-validation is an alternative to cross-validation and in the case of crystallographic refinement, the Hamilton test (Hamilton, 1965) is a prime example of this. This method enables one to assess whether a reduction in the *R* value is statistically significant given

the increase in the number of degrees of freedom. Application of this test in the case of macromolecules is compounded by the difficulty of estimating the effect of the combined set of restraints on the (effective) number of degrees of freedom, but some information can nevertheless be gained from such an analysis (Bacchi *et al.*, 1996).

21.1.7.4.2. *Real-space fits*

The fit of a model to the data can also be assessed in real space, which has the advantage that it can be performed for arbitrary sets of atoms (*e.g.* for every residue separately). Jones *et al.* (1991) introduced the real-space *R* value, which measures the similarity of a map calculated directly from the model (ρ_c) and one which incorporates experimental data (ρ_o) as

$$R = \sum |\rho_o - \rho_c| / \sum |\rho_o + \rho_c|,$$

where the sums extend over all grid points in the map that surround the selected set of atoms. The real-space fit can also be expressed as a correlation coefficient (Jones & Kjeldgaard, 1997), which has the advantage that no scaling of the two densities is necessary. Chapman (1995) described a modification in which the density calculated from the model is derived by Fourier transformation of resolution-truncated atomic scattering factors.

The program *SFCHECK* (Vaguine *et al.*, 1999) implements several variations on the real-space fit. The normalized average displacement measures the tendency of groups of atoms to move away from their current position. The density correlation is a modification of the real-space correlation coefficient. The residue-density index is calculated as the geometric mean of the density values of a set of atoms, divided by the average density of all atoms in the model. It therefore measures how high the electron-density level is for the set of atoms considered (*e.g.* all side-chain atoms of a residue). The connectivity index is identical to the residue-density index, but is calculated only for the N, C^{α} and C atoms. It thus provides an indication of the continuity of the main-chain electron density.

21.1.7.4.3. *Coordinate error estimates*

Since a measurement without an error estimate is not a measurement, crystallographers are keen to assess the estimated errors in the atomic coordinates and, by extension, in the atomic positions, bond lengths *etc.* In principle, upon convergence of a least-squares refinement, the variances and covariances of the model parameters (coordinates, ADPs and occupancies) may be obtained through inversion of the least-squares full matrix (Sheldrick, 1996; Ten Eyck, 1996; Cruickshank, 1999). In practice, however, this is seldom performed as the matrix inversion requires enormous computational resources. Therefore, one of a battery of (sometimes quasi-empirical) approximations is usually employed.

For a long time, the elegant method of Luzzati (1952) has been used for a different purpose (namely, to estimate average coordinate errors of macromolecular models) than that for which it was developed (namely, to estimate the positional changes required to reach a zero *R* value, using several assumptions that are not valid for macromolecules; Cruickshank, 1999). A Luzzati plot is a plot of *R* value *versus* $2 \sin\theta/\lambda$, and a comparison with theoretical curves is used to estimate the average positional error. Considering the problems with conventional *R* values (discussed in Section 21.1.7.4.1), Kleywegt *et al.* (1994) instead plotted free *R* values to obtain a cross-validated error estimate. This intuitive modification turned out to yield fairly reasonable values in practice (Kleywegt & Brünger, 1996; Brünger, 1997). Read (1986, 1990) estimated coordinate error from σ_A plots; the cross-validated modification of this method also yields reasonable error estimates (Brünger, 1997).

Cruickshank, almost 50 years after his work on the precision of small-molecule crystal structures (Cruickshank, 1949), introduced the diffraction-component precision index (DPI; Dodson *et al.*, 1996; Cruickshank, 1999) to estimate the coordinate or positional error of an atom with a *B* factor equal to the average *B* factor of the whole structure. In several cases for which full-matrix error estimates are available, the DPI gives quantitatively similar results. *SFCHECK* (Vaguine *et al.*, 1999) calculates both the DPI and Cruickshank's 1949 statistic (now termed the 'expected maximal error') based on the slope and the curvature of the electron-density map.

21.1.7.4.4. *Noncrystallographic symmetry*

Despite the multitude of criteria for assessing conformational differences between related molecules, there was until recently no objective way to assess whether such differences were a true reflection of the experimental data or a manifestation of refinement artifacts (Kleywegt & Jones, 1995*b*; Kleywegt, 1996). However, it has been found that electron-density maps calculated with experimental phases (or, at least, phases that are biased as little as possible by the model) and amplitudes can be used to correlate expected similarities (based on the data) with observed ones (manifest in the final refined models; Kleywegt, 1999). This method uses a local density-correlation map, as introduced by Read (Vellieux *et al.*, 1995), to measure the local similarity of the density of two or more models on a per-atom or per-residue basis. By comparing these values to the observed structural differences in the final models, it is relatively easy to check if the latter differences are warranted by the information contained in the experimental data (Kleywegt, 1999).

21.1.7.4.5. *Difference density quality*

van den Akker & Hol (1999) described a method (called *DDQ*, standing for difference density quality) to assess the local and global quality of a model based on analysis of an ($F_o - F_c$, α_c) map calculated after omission of all water molecules. In this method, the map and model are used to calculate several scores. One score assesses the presence or absence of favourably positioned water peaks near polar and apolar atoms. Other scores provide a measure for the presence or absence of positive and negative shift peaks that may indicate incorrect coordinates, temperature factors or occupancies. The scores can be averaged per residue or for an entire model and can be used to detect problems in models. The method appears to be applicable to $\sim$3 Å resolution.

21.1.7.5. *Accountancy*

The opinions as to what constitutes an error in a model vary somewhat in the community [compare Hooft *et al.* (1996) and Jones *et al.* (1996), for instance], but most people would agree that a crystallographic error is one that requires access to the experimental data for its verification, and whose correction alters the calculated structure factors (*e.g.* position, *B* factor, occupancy or scattering factors of one or more atoms). In addition to this, there are nomenclature rules and conventions to which a model that is made publicly available should adhere. Separate from this is the issue of the (more clerical) validation of public database entries ('PDB files'; Hooft *et al.*, 1994, 1996) which, while important to maintain the integrity of these databases, ultimately ought to be the responsibility of the database curators (Jones *et al.*, 1996; Keller *et al.*, 1998; Abola *et al.*, 2000).

21.1.8. Future

During the 1990s, the field of protein model validation matured rapidly (MacArthur *et al.*, 1994; EU 3-D Validation Network, 1998; Laskowski *et al.*, 1998) and further fundamental breakthroughs seem unlikely at present (although it would be highly desirable to be able to calculate and compare the information content of experimental data and models alike). In contrast, work on the validation of nucleic acid models (Schultze & Feigon, 1997) and hetero-entities (Kleywegt & Jones, 1998) has only just begun. In addition, there is still scope for further development of validation methods that use both the atomic model and the crystallographic data. In addition, the increasing number of structures that are solved at (near-)atomic resolution may lead to an adjustment of some validation criteria, *e.g.* of 'ideal' geometric target values, rotamer libraries *etc.* Also, validation of model aspects typically associated with very high resolution studies (refined occupancies, alternative conformations, anisotropic ADPs, H atoms) is still poorly developed. An increased understanding and appreciation of factors that determine model quality (and knowledge of how to measure them) will be important for the development of more automatic methods for protein structure determination. This in turn will enable 'black-box' high-throughput protein crystallography to become a reality, at least for 'run-of-the-mill' structures.

Acknowledgements

The author gratefully acknowledges the many useful discussions about validation with Alwyn Jones (Uppsala), Axel Brunger (Yale), Eleanor Dodson (York), Randy Read (Cambridge), Carl-Ivar Brändén (Stockholm), the members of the EU-funded 3-D Validation Network and many other colleagues in the field. This work was supported by the Swedish Foundation for Strategic Research (SSF), its Structural Biology Network (SBNet) and the EU-funded 3-D Validation Network.

References

21.1

Aalten, D. M. F. van, Bywater, R., Findlay, J. B. C., Hendlich, M., Hooft, R. W. W. & Vriend, G. (1996). *PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. J. Comput. Aided Mol. Des.* **10**, 255–262.

Abagyan, R. A. & Totrov, M. M. (1997). *Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. J. Mol. Biol.* **268**, 678–685.

Abola, E. E., Bairoch, A., Barker, W. C., Beck, S., Benson, D. A., Berman, H., Cantor, C., Doubet, S., Hubbard, T. J. P., Jones, T. A., Kleywegt, G. J., Kolaskar, A. S., van Kuik, A., Lesk, A. M., Mewes, H. W., Neuhaus, D., Pfeiffer, F., Ten Eyck, L. F., Simpson, R. J., Stoesser, G., Sussman, J. L., Tateno, Y., Tsugita, A., Ulrich, E. L. & Vliegenthart, J. F. G. (2000). *Quality control in databanks for molecular biology. BioEssays*, **22**, 1024–1034.

Adams, P. D., Pannu, N. S., Read, R. J. & Brünger, A. T. (1997). *Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. Proc. Natl Acad. Sci. USA*, **94**, 5018–5023.

Akker, F. van den & Hol, W. G. J. (1999). *Difference density quality (DDQ): a method to assess the global and local correctness of macromolecular crystal structures. Acta Cryst.* D**55**, 206–218.

Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). *The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. Acta Cryst.* B**35**, 2331–2339.

Allen, F. H. & Johnson, O. (1991). *Automated conformational analysis from crystallographic data. 4. Statistical descriptors for a distribution of torsion angles. Acta Cryst.* B**47**, 62–67.

Allen, F. H., Kennard, O., Watson, D. G., Brammer, L., Orpen, A. G. & Taylor, R. (1987). *Tables of bond lengths determined by X-ray and neutron diffraction. Part 1. Bond lengths in organic compounds. J. Chem. Soc. Perkin Trans.* 2, pp. S1–S19.

Ammon, H. L., Weber, I. T., Wlodawer, A., Harrison, R. W., Gilliland, G. L., Murphy, K. C., Sjölin, L. & Roberts, J. (1988). *Preliminary crystal structure of Acinetobacter glutaminasificans glutaminase-asparaginase. J. Biol. Chem.* **263**, 150–156.

Bacchi, A., Lamzin, V. S. & Wilson, K. S. (1996). *A self-validation technique for protein structure refinement: the extended Hamilton test. Acta Cryst.* D**52**, 641–646.

Balasubramanian, R. (1977). *New type of representation for mapping chain-folding in protein molecules. Nature (London)*, **266**, 856–857.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol.* **112**, 535–542.

Bhat, T. N. (1988). *Calculation of an OMIT map. J. Appl. Cryst.* **21**, 279–281.

Bhat, T. N. & Cohen, G. H. (1984). *OMITMAP: an electron density map suitable for the examination of errors in a macromolecular model. J. Appl. Cryst.* **17**, 244–248.

Borgstahl, G. E. O., Williams, D. R. & Getzoff, E. D. (1995). *1.4 Å structure of photoactive yellow protein, a cytosolic photoreceptor: unusual fold, active site, and chromophore. Biochemistry*, **34**, 6278–6287.

Bott, R. & Sarma, R. (1976). *Crystal structure of turkey egg-white lysozyme: results of the molecular replacement method at 5 Å resolution. J. Mol. Biol.* **106**, 1037–1046.

Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991). *A method to identify protein sequences that fold into a known three-dimensional structure. Science*, **253**, 164–170.

Brändén, C.-I. & Jones, T. A. (1990). *Between objectivity and subjectivity. Nature (London)*, **343**, 687–689.

Bricogne, G. & Irwin, J. (1996). *Maximum-likelihood refinement of incomplete models with BUSTER + TNT.* In *Proceedings of the CCP4 study weekend. Macromolecular refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.

Brünger, A. T. (1992*a*). *Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature (London)*, **355**, 472–475.

Brünger, A. T. (1992*b*). *X-PLOR. A system for crystallography and NMR.* Yale University, New Haven, Connecticut, USA.

Brünger, A. T. (1993). *Assessment of phase accuracy by cross validation: the free R value. Methods and applications. Acta Cryst.* D**49**, 24–36.

Brünger, A. T. (1997). *The free R value: a more objective statistic for crystallography. Methods Enzymol.* **277**, 366–396.

Carson, M., Buckner, T. W., Yang, Z., Narayana, S. V. L. & Bugg, C. E. (1994). *Error detection in crystallographic models. Acta Cryst.* D**50**, 900–909.

Carugo, O. & Bordo, D. (1999). *How many water molecules can be detected by protein crystallography? Acta Cryst.* D**55**, 479–483.

Chapman, M. S. (1995). *Restrained real-space macromolecular atomic refinement using a new resolution-dependent electron-density function. Acta Cryst.* A**51**, 69–80.

Colovos, C. & Yeates, T. O. (1993). *Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci.* **2**, 1511–1519.

Cruickshank, D. W. J. (1949). *The accuracy of electron-density maps in X-ray analysis with special reference to dibenzyl. Acta Cryst.* **2**, 65–82.

Cruickshank, D. W. J. (1950). *The convergence of the least-squares or Fourier refinement methods. Acta Cryst.* **3**, 10–13.

Cruickshank, D. W. J. (1999). *Remarks about protein structure precision. Acta Cryst.* D**55**, 583–601.

Diederichs, K. & Karplus, P. A. (1997). *Improved R-factors for diffraction data analysis in macromolecular crystallography. Nature Struct. Biol.* **4**, 269–275.

Dodson, E., Kleywegt, G. J. & Wilson, K. S. (1996). *Report of a workshop on the use of statistical validators in protein X-ray crystallography. Acta Cryst.* D**52**, 228–234.

Drenth, J. (1994). *Principles of protein X-ray crystallography.* New York: Springer–Verlag.

Engh, R. A. & Huber, R. (1991). *Accurate bond and angle parameters for X-ray protein structure refinement. Acta Cryst.* A**47**, 392–400.

EU 3-D Validation Network (1998). *Who checks the checkers? Four validation tools applied to eight atomic resolution structures. J. Mol. Biol.* **276**, 417–436.

Flocco, M. M. & Mowbray, S. L. (1995). *Cα-based torsion angles: a simple tool to analyze protein conformational changes. Protein Sci.* **4**, 2118–2122.

Greaves, R. B., Vagin, A. A. & Dodson, E. J. (1999). *Automated production of small-molecule dictionaries for use in crystallographic refinements. Acta Cryst.* D**55**, 1335–1339.

Gunasekaran, K., Ramakrishnan, C. & Balaram, P. (1996). *Disallowed Ramachandran conformations of amino acid residues in protein structures. J. Mol. Biol.* **264**, 191–198.

Hamilton, W. C. (1965). *Significance tests on the crystallographic R factor. Acta Cryst.* **18**, 502–510.

Hendrickson, W. A. (1985). *Stereochemically restrained refinement of macromolecular structures. Methods Enzymol.* **115**, 252–270.

Hendrickson, W. A. & Konnert, J. H. (1980). *Incorporation of stereochemical information into crystallographic refinement.* In *Computing in crystallography*, edited by R. Diamond, S. Ramaseshan & K. Venkatesan, pp. 13.01–13.25. Bangalore: Indian Academy of Science.

Herzberg, O. & Moult, J. (1991). *Analysis of the steric strain in the polypeptide backbone of protein molecules. Proteins Struct. Funct. Genet.* **11**, 223–229.

Hodel, A., Kim, S.-H. & Brünger, A. T. (1992). *Model bias in macromolecular crystal structures. Acta Cryst.* A**48**, 851–858.

21.1 (*cont.*)

Hoier, H., Schlömann, M., Hammer, A., Glusker, J. P., Carrell, H. L., Goldman, A., Stezowski, J. J. & Heinemann, U. (1994). *Crystal structure of chloromuconate cycloisomerase from Alcaligenes eutrophus JMP134 (pJP4) at 3 Å resolution. Acta Cryst.* D**50**, 75–84.

Hooft, R. W. W., Sander, C. & Vriend, G. (1994). *Reconstruction of symmetry-related molecules from Protein Data Bank (PDB) files. J. Appl. Cryst.* **27**, 1006–1009.

Hooft, R. W. W., Sander, C. & Vriend, G. (1996*a*). *Verification of protein structures: side-chain planarity. J. Appl. Cryst.* **29**, 714–716.

Hooft, R. W. W., Sander, C. & Vriend, G. (1996*b*). *Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. Proteins Struct. Funct. Genet.* **26**, 363–376.

Hooft, R. W. W., Sander, C. & Vriend, G. (1997). *Objectively judging the quality of a protein structure from a Ramachandran plot. Comput. Appl. Biosci.* **13**, 425–430.

Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Errors in protein structures. Nature (London)*, **381**, 272.

IUPAC–IUB Commission on Biochemical Nomenclature (1970). *Abbreviations and symbols for the description of the conformation of polypeptide chains. J. Mol. Biol.* **52**, 1–17.

James, M. N. G. & Sielecki, A. R. (1983). *Structure and refinement of penicillopepsin at 1.8 Å resolution. J. Mol. Biol.* **163**, 299–361.

Janin, J. (1990). *Errors in three dimensions. Biochimie*, **72**, 705–709.

Janin, J., Wodak, S., Levitt, M. & Maigret, B. (1978). *Conformation of amino acid side-chains in proteins. J. Mol. Biol.* **125**, 357–386.

Jones, T. A. & Kjeldgaard, M. (1994). *Making the first trace with O.* In *Proceedings of the CCP4 study weekend. From first map to final model*, edited by S. Bailey, R. Hubbard & D. A. Waller, pp. 1–13. Warrington: Daresbury Laboratory.

Jones, T. A. & Kjeldgaard, M. (1997). *Electron density map interpretation. Methods Enzymol.* **277**, 173–208.

Jones, T. A., Kleywegt, G. J. & Brünger, A. T. (1996). *Storing diffraction data. Nature (London)*, **381**, 18–19.

Jones, T. A. & Thirup, S. (1986). *Using known substructures in protein model building and crystallography. EMBO J.* **5**, 819–822.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Cryst.* A**47**, 110–119.

Karplus, P. A. (1996). *Experimentally observed conformation-dependent geometry and hidden strain in proteins. Protein Sci.* **5**, 1406–1420.

Keller, P. A., Henrick, K., McNeil, P., Moodie, S. & Barton, G. J. (1998). *Deposition of macromolecular structures. Acta Cryst.* D**54**, 1105–1108.

Kelly, J. A., Dideberg, O., Charlier, P., Wery, J. P., Libert, M., Moews, P. C., Knox, J. R., Duez, C., Fraipoint, C., Joris, B., Dusart, J., Frère, J. M. & Ghuysen, J. M. (1986). *On the origin of bacterial resistance to penicillin: comparison of a β-lactamase and a penicillin target. Science*, **231**, 1429–1431.

Kelly, J. A. & Kuzin, A. P. (1995). *The refined crystallographic structure of a DD-peptidase penicillin-target enzyme at 1.6 Å resolution. J. Mol. Biol.* **254**, 223–236.

Kleywegt, G. J. (1996). *Use of non-crystallographic symmetry in protein structure refinement. Acta Cryst.* D**52**, 842–857.

Kleywegt, G. J. (1997). *Validation of protein models from Cα coordinates alone. J. Mol. Biol.* **273**, 371–376.

Kleywegt, G. J. (1999). *Experimental assessment of differences between related protein crystal structures. Acta Cryst.* D**55**, 1878–1884.

Kleywegt, G. J., Bergfors, T., Senn, H., Le Motte, P., Gsell, B., Shudo, K. & Jones, T. A. (1994). *Crystal structures of cellular retinoic acid binding proteins I and II in complex with all-trans-retinoic acid and a synthetic retinoid. Structure*, **2**, 1241–1258.

Kleywegt, G. J. & Brünger, A. T. (1996). *Checking your imagination: applications of the free R value. Structure*, **4**, 897–904.

Kleywegt, G. J., Hoier, H. & Jones, T. A. (1996). *A re-evaluation of the crystal structure of chloromuconate cycloisomerase. Acta Cryst.* D**52**, 858–863.

Kleywegt, G. J. & Jones, T. A. (1995*a*). *Braille for pugilists.* In *Proceedings of the CCP4 study weekend. Making the most of your model*, edited by W. N. Hunter, J. M. Thornton & S. Bailey, pp. 11–24. Warrington: Daresbury Laboratory.

Kleywegt, G. J. & Jones, T. A. (1995*b*). *Where freedom is given, liberties are taken. Structure*, **3**, 535–540.

Kleywegt, G. J. & Jones, T. A. (1996*a*). *Efficient rebuilding of protein structures. Acta Cryst.* D**52**, 829–832.

Kleywegt, G. J. & Jones, T. A. (1996*b*). *Phi/Psi-chology: Ramachandran revisited. Structure*, **4**, 1395–1400.

Kleywegt, G. J. & Jones, T. A. (1997). *Model-building and refinement practice. Methods Enzymol.* **277**, 208–230.

Kleywegt, G. J. & Jones, T. A. (1998). *Databases in protein crystallography. Acta Cryst.* D**54**, 1119–1131.

Kleywegt, G. J. & Read, R. J. (1997). *Not your average density. Structure*, **5**, 1557–1569.

Kleywegt, G. J., Zou, J. Y., Divne, C., Davies, G. J., Sinning, I., Ståhlberg, J., Reinikainen, T., Srisodsuk, M., Teeri, T. T. & Jones, T. A. (1997). *The crystal structure of the catalytic core domain of endoglucanase I from Trichoderma reesei at 3.6 Å resolution, and a comparison with related enzymes. J. Mol. Biol.* **272**, 383–397.

Korn, A. P. & Rose, D. R. (1994). *Torsion angle differences as a means of pinpointing local polypeptide chain trajectory changes for identical proteins in different conformational states. Protein Eng.* **7**, 961–967.

Kuriyan, J. & Weis, W. I. (1991). *Rigid protein motion as a model for crystallographic temperature factors. Proc. Natl Acad. Sci. USA*, **88**, 2773–2777.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst.* **26**, 283–291.

Laskowski, R. A., MacArthur, M. W. & Thornton, J. M. (1994). *Evaluation of protein coordinate data sets.* In *Proceedings of the CCP4 study weekend. From first map to final model*, edited by S. Bailey, R. Hubbard & D. A. Waller, pp. 149–159. Warrington: Daresbury Laboratory.

Laskowski, R. A., MacArthur, M. W. & Thornton, J. M. (1998). *Validation of protein models derived from experiment. Curr. Opin. Struct. Biol.* **8**, 631–639.

Laskowski, R. A., Moss, D. S. & Thornton, J. M. (1993). *Main-chain bond lengths and bond angles in protein structures. J. Mol. Biol.* **231**, 1049–1067.

Lubkowski, J., Wlodawer, A., Housset, D., Weber, I. T., Ammon, H. L., Murphy, K. C. & Swain, A. L. (1994). *Refined crystal structure of Acinetobacter glutaminasificans glutaminase-asparaginase. Acta Cryst.* D**50**, 826–832.

Lüthy, R., Bowie, J. U. & Eisenberg, D. (1992). *Assessment of protein models with three-dimensional profiles. Nature (London)*, **356**, 83–85.

Luzzati, V. (1952). *Traitement statistique des erreurs dans la determination des structures crystallines. Acta Cryst.* **5**, 802–810.

MacArthur, M. W., Laskowski, R. A. & Thornton, J. M. (1994). *Knowledge-based validation of protein structure coordinates derived by X-ray crystallography and NMR spectroscopy. Curr. Opin. Struct. Biol.* **4**, 731–737.

MacArthur, M. W. & Thornton, J. M. (1996). *Deviations from planarity of the peptide bond in peptides and proteins. J. Mol. Biol.* **264**, 1180–1195.

MacArthur, M. W. & Thornton, J. M. (1999). *Protein side-chain conformation: a systematic variation of χ_1 mean values with resolution – a consequence of multiple rotameric states? Acta Cryst.* D**55**, 994–1004.

McDonald, I. K. & Thornton, J. M. (1995). *The application of hydrogen bonding analysis in X-ray crystallography to help orientate asparagine, glutamine and histidine side chains. Protein Eng.* **8**, 217–224.

McRee, D. E., Tainer, J. A., Meyer, T. E., van Beeumen, J., Cusanovich, M. A. & Getzoff, E. D. (1989). *Crystallographic*

21.1 (*cont.*)

structure of a photoreceptor protein at 2.4 Å resolution. Proc. Natl Acad. Sci. USA, **86**, 6533–6537.
Maiorov, V. & Abagyan, R. (1998). *Energy strain in three-dimensional protein structures. Fold. Des.* **3**, 259–269.
Marsh, R. E. (1995). *Some thoughts on choosing the correct space group. Acta Cryst.* B**51**, 897–907.
Marsh, R. E. (1997). *The perils of Cc revisited. Acta Cryst.* B**53**, 317–322.
Melo, F. & Feytmans, E. (1998). *Assessing protein structures with a non-local atomic interaction energy. J. Mol. Biol.* **277**, 1141–1152.
Merritt, E. A. (1999). *Expanding the model: anisotropic displacement parameters in protein structure refinement. Acta Cryst.* D**55**, 1109–1117.
Merritt, E. A., Kuhn, P., Sarfaty, S., Erbe, J. L., Holmes, R. K. & Hol, W. G. J. (1998). *The 1.25 Å resolution refinement of the cholera toxin B-pentamer: evidence of peptide backbone strain at the receptor-binding site. J. Mol. Biol.* **282**, 1043–1059.
Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). *Stereochemical quality of protein structure coordinates. Proteins Struct. Funct. Genet.* **12**, 345–364.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Refinement of macromolecular structures by the maximum-likelihood method. Acta Cryst.* D**53**, 240–255.
Nayal, M. & Di Cera, E. (1996). *Valence screening of water in protein crystals reveals potential Na^+ binding sites. J. Mol. Biol.* **256**, 228–234.
Noble, M. E. M., Zeelen, J. P., Wierenga, R. K., Mainfroid, V., Goraj, K., Gohimont, A. C. & Martial, J. A. (1993). *Structure of triosephosphate isomerase from Escherichia coli determined at 2.6 Å resolution. Acta Cryst.* D**49**, 403–417.
Nunn, R. S., Artymiuk, P. J., Baker, P. J., Rice, D. W. & Hunter, C. N. (1995). *Retraction – Purification and crystallization of the light harvesting LH1 complex from Rhodobacter sphaeroides. J. Mol. Biol.* **252**, 153.
Oldfield, T. J. & Hubbard, R. E. (1994). *Analysis of Cα geometry in protein structures. Proteins Struct. Funct. Genet.* **18**, 324–337.
Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Incorporation of prior phase information strengthens maximum-likelihood structure refinement. Acta Cryst.* D**54**, 1285–1294.
Pannu, N. S. & Read, R. J. (1996). *Improved structure refinement through maximum likelihood. Acta Cryst.* A**52**, 659–668.
Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A. T. & Berman, H. M. (1996). *New parameters for the refinement of nucleic acid-containing structures. Acta Cryst.* D**52**, 57–64.
Ponder, J. W. & Richards, F. M. (1987). *Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol.* **193**, 775–791.
Pontius, J., Richelle, J. & Wodak, S. J. (1996). *Deviations from standard atomic volumes as a quality measure for protein crystal structures. J. Mol. Biol.* **264**, 121–136.
Priestle, J. P. (1994). *Stereochemical dictionaries for protein structure refinement and model building. Structure*, **2**, 911–913.
Ramachandran, G. N. & Mitra, A. K. (1976). *An explanation for the rare occurrence of cis peptide units in proteins and polypeptides. J. Mol. Biol.* **107**, 85–92.
Ramachandran, G. N. & Srinivasan, R. (1961). *An apparent paradox in crystal structure analysis. Nature (London)*, **190**, 159–161.
Ramakrishnan, C. & Ramachandran, G. N. (1965). *Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units. Biophys. J.* **5**, 909–933.
Read, R. J. (1986). *Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Cryst.* A**42**, 140–149.
Read, R. J. (1990). *Structure-factor probabilities for related structures. Acta Cryst.* A**46**, 900–912.
Read, R. J. (1994). *Model bias and phase combination.* In *Proceedings of the CCP4 study weekend. From first map to final model*, edited by S. Bailey, R. Hubbard & D. A. Waller, pp. 31–40. Warrington: Daresbury Laboratory.
Read, R. J. (1997). *Model phases: probabilities and bias. Methods Enzymol.* **277**, 110–128.
Schomaker, V. & Trueblood, K. N. (1968). *On the rigid-body motion of molecules in crystals. Acta Cryst.* B**24**, 63–76.
Schultze, P. & Feigon, R. (1997). *Chirality errors in nucleic acid structures. Nature (London)*, **387**, 668.
Sevcik, J., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1996). *Ribonuclease from Streptomyces aureofaciens at atomic resolution. Acta Cryst.* D**52**, 327–344.
Sheldrick, G. M. (1996). *Least-squares refinement of macromolecules: estimated standard deviations, NCS restraints and factors affecting convergence.* In *Proceedings of the CCP4 study weekend. Macromolecular refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 47–58. Warrington: Daresbury Laboratory.
Sinning, I., Kleywegt, G. J., Cowan, S. W., Reinemer, P., Dirr, H. W., Huber, R., Gilliland, G. L., Armstrong, R. N., Ji, X., Board, P. G., Olin, B., Mannervik, B. & Jones, T. A. (1993). *Structure determination and refinement of human alpha class glutathione transferase A1-1, and a comparison with the mu and pi class enzymes. J. Mol. Biol.* **232**, 192–212.
Sippl, M. J. (1993). *Recognition of errors in three-dimensional structures of proteins. Proteins Struct. Funct. Genet.* **17**, 355–362.
Stewart, D. E., Sarkar, A. & Wampler, J. E. (1990). *Occurrence and role of cis peptide bonds in protein structures. J. Mol. Biol.* **214**, 253–260.
Swindells, M. B., MacArthur, M. W. & Thornton, J. M. (1995). *Intrinsic φ, ψ propensities of amino acids, derived from the coil regions of known structures. Nature Struct. Biol.* **2**, 596–603.
Taylor, R. & Kennard, O. (1986). *Accuracy of crystal structure error estimates. Acta Cryst.* B**42**, 112–120.
Ten Eyck, L. F. (1996). *Full matrix least squares.* In *Proceedings of the CCP4 study weekend. Macromolecular refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 37–45. Warrington: Daresbury Laboratory.
Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998). *R_{free} and the R_{free} ratio. I. Derivation of expected values of cross-validation residuals used in macromolecular least-squares refinement. Acta Cryst.* D**54**, 547–557.
Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Cryst.* D**55**, 191–205.
Vellieux, F. M. D. A. P., Hunt, J. F., Roy, S. & Read, R. J. (1995). *DEMON/ANGEL: a suite of programs to carry out density modification. J. Appl. Cryst.* **28**, 347–351.
Vriend, G. (1990). *WHAT IF: a molecular modeling and drug design program. J. Mol. Graphics*, **8**, 52–56.
Vriend, G. & Sander, C. (1993). *Quality control of protein models: directional atomic contact analysis. J. Appl. Cryst.* **26**, 47–60.
Walther, D. & Cohen, F. E. (1999). *Conformational attractors on the Ramachandran map. Acta Cryst.* D**55**, 506–517.
Watkin, D. (1996). *Pseudo symmetry.* In *Proceedings of the CCP4 study weekend. Macromolecular refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 171–184. Warrington: Daresbury Laboratory.
Weaver, L. H., Tronrud, D. E., Nicholson, H. & Matthews, B. W. (1990). *Some uses of the Ramachandran (φ, ψ) diagram in the structural analysis of lysozymes. Curr. Sci.* **59**, 833–837.
Weiss, M. S. & Hilgenfeld, R. (1997). *On the use of the merging R factor as a quality indicator for X-ray data. J. Appl. Cryst.* **30**, 203–205.
Wilson, A. J. C. (1949). *The probability distribution of X-ray intensities. Acta Cryst.* **2**, 318–321.
Zou, J. Y. & Mowbray, S. L. (1994). *An evaluation of the use of databases in protein structure refinement. Acta Cryst.* D**50**, 237–249.

21.2

Abagyan, R. A. & Totrov, M. M. (1997). *Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. J. Mol. Biol.* **268**, 678–685.