International Tables for Crystallography (2006). Vol. F, Section 21.1.2, pp. 497–498.

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

references