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