21.3. Detection of errors in protein models

By O. Dym, D. Eisenberg and T. O. Yeates

21.3.1. Motivation and introduction

The discovery of major errors in several protein structural models determined by X-ray crystallography has focused attention on methods of detecting and minimizing such errors. There are several sources of error in the determination of a protein structure. Errors enter not only in the collection of the experimental data, but especially in their interpretation. Limited diffraction resolution and poor phases frequently lead to electron-density maps that are difficult to interpret. As a result, preliminary protein models built into ambiguous maps often contain errors of various types. The different types of errors can be arranged in decreasing order of severity, as follows: mistracing of the protein chain due to uncertainty in backbone connectivity, misalignment or misregistration of residues, and misplacement of side-chain and backbone atoms. It is critical to be able to identify these problematic regions of a model so they can be given special attention during the iterative process of model building and atomic refinement.

During atomic refinement, the atomic coordinates of the macromolecule are adjusted to minimize an error function of two terms. The first term contains the discrepancies between the observed diffraction data and structure factors calculated from the model. The second term describes the deviations from ideal geometry, such as deviations in bond lengths, bond angles, planarity and other specific features. When refinement is complete, the residual errors in the separate terms are reported, with the discrepancies in the diffraction data embodied in the R value. These error values are usually taken as the first indicators of structure quality.

Beyond criteria that are explicitly minimized during refinement, other structural properties may be devised and evaluated. Some properties that have been investigated include the distribution of non-polar and polar residues both on the surface and in the interior of the protein, and preferred environments for different atom types and residues. These measures use the empirical knowledge gathered in the Protein Data Bank (PDB) to assess how 'normal' or 'abnormal' a given model is. The measures are also useful in cases in which the experimental diffraction data are not available (*e.g.* when assessing structures already in the data bank). Several programs that validate protein structural models on the basis of various structural properties are available. Among them are *PROCHECK* (Laskowski *et al.*, 1993), *WHAT IF* (Vriend, 1990; Vriend & Sander, 1993), *ERRAT* (Colovos & Yeates, 1993), and *VERIFY*3*D* (Lüthy *et al.*, 1992; Bowie *et al.*, 1991). The various programs have the same objectives, but differ in many important respects. The approaches differ with regard to the scale of the analysis (*e.g.* atom-based *versus* amino-acid based), the level of detail in the program output, and the degree to which the evaluated properties are independent of the refinement function.

21.3.2. Separating evaluation from refinement

Any property that has been constrained or heavily restrained during refinement of the atomic model, and any property that has been closely monitored during rebuilding, cannot be used as the sole criterion to assess or 'prove' the quality of the model. The reason is that if the atomic model is adjusted to optimize a particular property, that property no longer gives an unbiased measure of model accuracy. For example, most refinement programs operate by adjusting atomic positions to minimize the difference between observed and calculated structure-factor amplitudes, known as the R factor or R value. Since the R value is the target of the optimization procedure, it does not provide an *independent* measure of quality. As a result, the ordinary R value can be misleading. A much more reliable measure is the free R value (Brünger, 1992), which is calculated from a randomly selected subset of the diffraction data that are excluded from the atomic refinement calculations. The importance of using the free R value to monitor refinement is now widely accepted.

Likewise, independent criteria must be employed to judge protein models themselves, aside from the diffraction data. Typical atomic refinement protocols tightly restrain the obvious stereochemical terms, such as bond lengths, angles and planarity. Therefore, low deviation from ideal geometry cannot be presented as proof of the quality of the structure. Independent criteria must be based on higher-level geometric considerations. Several programs that include such evaluations are described here.

Criteria that are useful for assessing the validity of protein models are those that are not directly restrained during the process of refinement. The following three properties of protein models are of this type: (1) the main-chain dihedral angles; (2) the non-bonded interactions of protein atoms with other protein atoms and with the solvent; and (3) the packing of atoms within the structure. Each of these properties of a proposed model can be compared for consistency with the same property observed in a database of trustworthy structures. To the extent that the property deviates from the values observed for the proteins of the database, the proposed model is suspect. Some of these properties can be computed for each segment of a protein or for local regions in three-dimensional (3D) space. In this way, inaccurate regions within a proposed model can be identified.

21.3.3. Algorithms for the detection of errors in protein models and the types of errors they detect

21.3.3.1. *PROCHECK*

The *PROCHECK* (Laskowski *et al.*, 1993) suite of programs compares the stereochemistry of a proposed protein model to stereochemical features of known structures. The program provides an assessment of the overall quality of the model by comparing the model with well refined structures of the same resolution, and also highlights regions that may need further adjustment. The output of *PROCHECK* comprises a number of plots, together with detailed residue-by-residue listings of secondary-structure assignment, non-bonded interactions between different pairs of residues, main-chain bond lengths and bond angles, and peptide-bond planarity.

The program also displays main-chain dihedral angles (φ and ψ) as a two-dimensional Ramachandran (Ramachandran & Sasisekharan, 1968) plot. The Ramachandran plot classifies each residue in one of three categories: 'allowed' conformations; 'partially allowed' conformations, which give rise to modestly unfavourable repulsion between non-bonded atoms, and which might be overcome by attractive effects such as hydrogen bonds; and 'disallowed' interactions which give highly unfavourable non-bonded interatomic distances. The Ramachandran plot can identify unacceptable clusters of φ–ψ angles, revealing possible errors made during model building and refinement. As opposed to covalent bond angles and bond lengths, the main-chain dihedral angles are not usually restrained during X-ray refinement and therefore can be used to validate the structural model independently. In practice, the Ramachandran plot is one of the simplest, most sensitive tools for assessing the quality of a protein model.