International Tables for Crystallography (2006). Vol. F, Section 21.3.3, pp. 520–521.

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 *VERIFY3D* (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.

The *PROCHECK* suite is generally useful for assessing the quality of protein structures in various stages of completion. The Ramachandran analysis is especially informative. However, it is possible, at least in principle, to devise an incorrect model with fully acceptable main-chain and side-chain stereochemistry, so other methods must also be used to assess protein models.

21.3.3.2. *WHAT IF*

The molecular modelling and drug design program *WHAT IF* (Vriend, 1990) performs a large number of geometrical checks, comparing a proposed protein model to a set of canonical distances and angles. These parameters include bond lengths and bond angles, side-chain planarity, torsion angles, interatomic distances, unusual backbone conformations and the Ramachandran plot. New additions (Vriend & Sander, 1993) include a 'quality factor', and a number of checks for clashes between symmetry-related molecules. Starting from the hypothesis that atom–atom interactions are the primary determinant of protein folding, the program tests a protein model for proper packing by calculating a contact quality index. Each contact is characterized by its fragment type (80 types from the 20 residues), the atom type and the three-dimensional location of the atom relative to the local frame of the fragment. Sets of database-derived distributions are compared with the actual distribution in the protein model being tested. A good agreement with the database distribution produces a high contact quality index. A low packing score can indicate any of: poor packing, misthreading of the sequence, bad crystal contacts, bad contacts with a co-factor, or proximity to a vacant active site. The contact analysis available in *WHAT IF* can be used as an independent quality indicator during crystallographic refinement, or during the process of protein modelling and design.

21.3.3.3. *VERIFY3D*

The program *VERIFY3D* (Lüthy *et al.*, 1992; Bowie *et al.*, 1991) measures the compatibility of a protein model with its own amino-acid sequence. Each residue position in the 3D model is characterized by its environment and is represented by a row of 20 numbers in a '3D profile'. These numbers are the statistical preferences, called 3D–1D scores, of each of the 20 amino acids for this environment. Environments of residues are defined by three parameters: the area of the residue buried in the protein and inaccessible to solvent, the fraction of the side-chain area that is covered by polar atoms (O and N), and the local secondary structure. The 3D profile score, *S*, for the compatibility of the sequence with the model is the sum, over all residue positions, of the 3D–1D scores for the amino-acid sequence of the protein. The compatibility of segments of the sequence with their 3D structures can be assessed by plotting, against sequence number, the average 3D–1D score in a window of 21 residues. The 3D profile method rests on the observation that soluble proteins bury many hydrophobic side chains and not many polar residues.

Three applications for 3D profiles exist. The first is to assess the validity of protein models (Lüthy *et al.*, 1992). For 3D protein models known to be correct, the 3D profile score, *S*, for the compatibility of the amino-acid sequence with the environments formed by the model is high. In contrast, *S* for the compatibility with its sequence of a totally or partially wrong 3D protein model is generally low. Therefore, models that are largely incorrect or models that contain a small number of improperly built segments can be detected by low-scoring regions in the 3D profile. However, not all faulty regions are always evident directly from the profile, particularly if the misbuilt regions are at the termini, where they are obscured by the windowing procedure. The second application is to assess which is the stable oligomeric state of the folded protein, by comparing the accessibility (buried or exposed) of amino-acid side chains in the monomeric and oligomeric state (Eisenberg *et al.*, 1992). The third application is to identify other protein sequences which are folded in the same general pattern as the structure from which the profile was prepared (Bowie *et al.*, 1991). Predicting a protein structure from sequence requires a link between 3D structure and 1D sequence. The program *VERIFY*3*D* provides this link by reducing a 3D structure to 1D string of environmental classes. Therefore the method can be used to evaluate any protein model or to measure the compatibility of any protein structure with its amino-acid sequence.

21.3.3.4. *ERRAT*

The program *ERRAT* (Colovos & Yeates, 1993) analyses the relative frequencies of noncovalent interactions between atoms of various types. It can be viewed as an extension of the earlier 3D profile approach from the residue level to the atom level. Three types of atoms are considered (C, N and O), and consequently six types of interactions are possible (CC, CN, CO, NN, NO and OO).

ERRAT operates under the hypothesis that different atom types will be distributed non-randomly with respect to each other in proteins due to complex geometric and energetic considerations, and that structural errors will lead to detectable anomalies in the pattern of interactions. Assessment of the non-bonded interactions is subject to the following restrictions: the distance between the two atoms in space is less than some preset limit, typically 3.5 Å, and the atoms within the same residue or those that are covalently bonded to each other are not considered. For each nine-residue segment of sequence, the non-bonded contacts to other atoms in the protein are tallied by atomic interaction type and the result is divided by the total number of interactions. This gives a list, or six-dimensional vector, of fractional interaction frequencies that add up to unity. In this way, each nine-residue fragment generates one point in a five-dimensional space; only five of the six fractional values are independent. A large number of observations were extracted from reliable high-resolution structures and used to establish a multi-variate five-dimensional normal distribution for accurate protein structures. This distribution is used to evaluate the probability that a given set of interactions from a protein model in question is correct. Since the *ERRAT* evaluation is based on a normal distribution calibrated on a reliable database, it is straightforward to estimate the likelihood that each region of a candidate protein model is incorrect. This method provides an unbiased and statistically sound tool for identifying incorrectly built regions in protein models.

21.3.4. Selection of database

Regardless of the specific approach or the specific criteria for validating structural models, a reliable reference database has to be chosen by careful selection of known structures. Suitable criteria to consider when selecting a database are: protein structures determined to resolutions of 2.5 Å or better, *R* factors less than 25%, and good geometry, particularly of the dihedral angles of the protein backbone. In addition, the database should include examples from many diverse classes of structures and at the same time avoid multiple identical structures.

21.3.5. Examples: detection of errors in structures

21.3.5.1. *Specific examples*

Several examples are presented of errors in structural models determined by X-ray crystallography that can be detected using validation methods. One is that of the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO), which was traced

references