21.2. Assessing the quality of macromolecular structures

BY S. J. WODAK, A. A. VAGIN, J. RICHELLE, U. DAS, J. PONTIUS AND H. M. BERMAN

21.2.1. Introduction

X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are the two major techniques that provide detailed information on the atomic structure of macromolecules. Usually, however, the data obtained from these techniques are not of high enough resolution to define the atomic positions of a macromolecule with sufficient precision. Deriving the atomic models from the experimental data therefore involves sophisticated optimization (refinement) procedures, in which constraints based on prior knowledge about the chemical structure of the molecule and its conformational properties are applied. The resulting models are therefore prone to errors, which fall into two broad categories: systematic errors caused by biases during the structure determination and refinement procedures, and random errors which affect the precision of the model. Moreover, the quality of the model can vary in different regions of the structure, often due to higher local conformational or thermal disorder in certain parts.

With the rapid growth in the number of structures of macromolecules determined and the spreading use of structural information in different areas of science, the availability of objective criteria and methods for evaluating the quality of these structures has become a very important requirement.

A variety of validation procedures have been proposed by many groups (for recent reviews, see MacArthur *et al.*, 1994, and Laskowski *et al.*, 1998). The procedures involve two main approaches. One approach comprises procedures that validate the geometric and conformational parameters of the final model. This is done by measuring the extent to which the parameters deviate from standard values, derived from crystals of small molecules or from a set of high-quality structures of other macromolecules. The main limitation of this approach is that the quality of a model is defined by comparison with other known models without taking into account the experimental data. This harbours the danger of considering unusual conformations related to biological function as errors in the model or of accepting as 'normal' only what has already been observed before. The second and most important approach by far comprises procedures that take into account the experimental data and evaluate the agreement of the atomic model with these data. These procedures can, in principle, evaluate systematic errors and biases that affect the global quality of the model and can also detect local imprecision. The most commonly cited measures of agreement between the model as a whole and the data are the R factor and the 'free R factor', $R_{\rm free}$. Criteria for evaluating the local agreement of the model with electron density on a per-atom or per-residue basis are also available, and, more recently, access to more powerful computers has made it possible to compute the standard uncertainties of individual parameters, such as atomic coordinates or thermal factors.

Finally, the growing number of atomic resolution structures – primarily of proteins – is starting to provide a valuable source of much more precise information about the structures' geometrical and conformational properties. This should contribute to the improvement of standard values used in validation.

In this chapter, we present an overview of the different types of validation procedures applied to proteins and nucleic acids. We illustrate, in some detail, an approach to model validation based on atomic volumes embodied in the program *PROVE* and describe the package *SFCHECK*, which combines a set of criteria for evaluating the quality of the experimental data and the agreement of the model with the data.

21.2.2. Validating the geometric and stereochemical parameters of the model

21.2.2.1. *Comparisons against standard values derived from crystals of small molecules*

This concerns the validation of the covalent geometry of the atomic model. It involves comparing the bond distances and angles of the macromolecule against standard values and their associated uncertainties, derived from crystal structures of small organic molecules available in the Cambridge Structural Database, CSD (Allen *et al.*, 1979, 1983).

The standard values derived in this way are also used as restraints in crystallographic refinement programs, such as *XPLOR* (Brünger, 1992*a*) or the *CCP*4 suite of programs (Collaborative Computational Project, Number 4, 1994). As a result, the bond distances and angles of the final model usually agree well with their standard values, and the degree of scatter merely reflects the relative weight imposed on the various terms of the target function during refinement.

For proteins, the most commonly used standard values for the bond distances and angles are those compiled by Engh & Huber (1991) from molecular fragments in the CSD that most closely resemble chemical groups in amino acids. These parameters were shown to yield an improved description over that provided by the param19x.pro used in *XPLOR*, especially for the covalent geometry of aromatic rings in side-chain groups. It is noteworthy that these CSD-derived bond distances and angles can differ significantly from those used in molecular dynamics force fields, such as that of a recent version of *CHARMM* (MacKerell *et al.*, 1998). In these force fields, covalent-geometry parameters are obtained by a different strategy. They are optimized together with non-bonded parameters against a large body of available energy and structural data for a limited set of compounds representing amino-acid building blocks.

Protein-structure validation packages, such as *PROCHECK* (Laskowski *et al.*, 1993) and *WHAT IF* (Hooft, Vriend *et al.*, 1996), flag all bond distances and angles that deviate significantly from the database-derived reference values. This includes analysis of the deviations from planarity in aromatic rings and planar side-chain groups.

Similar checks are performed for the covalent geometry of atomic models of RNA or DNA oligo- and polynucleotides. Here, standard ranges for bond distances and angles are derived from crystal structures of nucleic acid bases, mononucleosides and mononucleotides in the CSD (Clowney *et al.*, 1996; Gelbin *et al.*, 1996). These values are used in validation procedures developed by the Nucleic Acid Database (NDB) (Berman *et al.*, 1992) and in crystallographic refinement programs. For higher-resolution structures (better than 2.4 Å), a standard geometry, dependent on the sugar pucker conformation (C2′ *endo* or C3′ *endo*) (Parkinson *et al.*, 1996), is used.

Validation of the covalent geometry of the so-called 'hetero groups' (chemically modified monomer groups or small molecules that bind to macromolecules) is much more difficult. It therefore tends not to be routinely performed, and, as a result, the quality of the hetero groups in the models deposited in the Protein Data Bank (PDB) (Bernstein *et al.*, 1977; Berman *et al.*, 2000) varies widely.

The variety of the chemical structures of these molecules (the current release of the PDB contains about 2700 chemically distinct compounds) makes it difficult to archive them consistently, let alone to compile the dictionaries containing the required reference

references