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 *VERIFY3D* 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 multivariate 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