International Tables for Crystallography (2006). Vol. F, Section 21.2.2.2.2, pp. 508–509.

values in advance. Proper handling and verification of such groups require a comprehensive and rigorous description of the chemical components, as well as flexible means of deriving the appropriate reference geometries.

The development of systematic procedures for checking bond lengths and various torsion angles of hetero groups (Kleywegt & Jones, 1998) is a step in the right direction. Further progress should come, thanks in part to the recently adopted macromolecular Crystallographic Information File (mmCIF) format (Bourne *et al.*, 1997), which provides the necessary framework for a much more comprehensive and rigorous description of the molecular components. Using this description as the basis, automated tools for building 'customized' dictionaries of geometrical standards have been developed. One such tool is A LigAnd and Monomer Object Data Environment (*A LA MODE*) (Clowney *et al.*, 1999). It starts from a minimal topological description of a ligand or monomer component and performs the tasks required to construct the mmCIF component description. This includes querying the CSD, integration and book-keeping of database survey results, analysis and comparison of covalent geometry and stereochemistry, and the assembly of complex model structures from the results of multiple database surveys. Tools such as this considerably simplify the handling of small molecules at the refinement, validation and archiving stages.

21.2.2.2. *Comparisons against standard values derived from surveys of other macromolecules*

This involves computing a number of stereochemical, geometric and energy parameters from the atomic coordinates of the macromolecule and comparing them with standard ranges derived from high-quality crystal structures of other macromoleules. These standards represent the 'expected' properties, and the aim is to evaluate the quality of a model by measuring the extent to which it departs from these properties.

This evaluation is usually performed at the global level, in order to assess the quality of the structure as a whole, and on the local level, to identify specific regions with unusual properties. Such regions may represent genuine problems with the model or unusual conformations adopted for functional purposes, and it is sometimes difficult to distinguish between these two alternatives. The choice of reference structures from which the standards are derived is a crucial aspect of the approach, since both the mean and shape of the reference distributions may be affected by it.

21.2.2.2.1. *Validation of stereochemical and non-bonded parameters*

Morris *et al.* (1992) pioneered this type of validation for proteins. The software *PROCHECK* (Laskowski *et al.*, 1993), which implements and extends this approach, is described in detail in Chapter 25.2 of this volume. A very important evaluation criterion is the Ramachandran-plot quality, where the distribution of the backbone φ, ψ angles of a given protein structure is compared to that in high-quality structures. The comparison is performed both globally, by determining the proportion of the residues in favourable (core) regions of the plot, and locally, by the log-odds (*G*-factor) value, which measures how normal or unusual a residue's location is in the plot for a given residue type.

A similar strategy is used to evaluate other stereochemical parameters, such as the side-chain torsion angles (χ_1, χ_2, χ_3 *etc.*), the peptide bond torsion (ω), the C$^\alpha$ tetrahedral distortion, disulfide bond geometry and stereochemistry.

An evaluation of the backbone hydrogen-bonding energy is also performed, using the Kabsch & Sander (1983) algorithm, by comparison with distributions computed from high-resolution protein structures.

Other programs like *WHAT IF* (Hooft, Vriend *et al.*, 1996) perform similar evaluations. This program computes the expected φ, ψ distribution for each residue type from a data set of non-redundant high-quality structures and evaluates how the φ, ψ distribution of a given protein deviates from the expected values (Hooft *et al.*, 1997). A somewhat different version of this approach is proposed by Kleywegt & Jones (1996). *WHAT IF* also computes other quality indicators such as the number of buried unsatisfied hydrogen bonds or the extent of the overlap of van der Waals spheres ('clashes'). In addition, it verifies the orientation of His, Gln and Asn side chains, based on a hydrogen-bond network analysis, which also takes into account hydrogen bonds between symmetry-related molecules (Hooft, Sander & Vriend, 1996).

The very small fraction of structures ($< 1.3\%$) for which only the C$^\alpha$ coordinates are deposited cannot be validated by the standard techniques. For these structures, two sets of parameters were shown to be useful (Kleywegt & Jones, 1996). They are the C$^\alpha$—C$^\alpha$ distances and a Ramachandran-like plot which displays for each residue the C^α_{i-1}—C^α_i—C^α_{i+1}—C^α_{i+2} dihedral angle against the C^α_{i-1}—C^α_i—C^α_{i+1} angle. Deviations from the expected distributions of these parameters, computed from a set of high-quality complete protein structures, are used as quality indicators.

The validation of nucleic acid stereochemistry, in particular DNA, has a much shorter history. Only in recent years has the number of high-quality nucleic acid crystal structures become large enough to permit the derivation of reliable conformational trends. Schneider *et al.* (1997) derived ranges and mean values for the torsion angles of the sugar–phosphate backbone in helical DNA from a set of 96 oligodeoxynucleotide crystal structures. These ranges form the basis for the nucleic acid structure validation protocols currently implemented at the NDB.

21.2.2.2.2. *Validation using knowledge-based interaction potentials and profiles*

These methods represent a distinct set of approaches to the validation of the non-bonded and conformational parameters of the model. They involve computing the relative frequencies of residue–residue or atom–atom contacts from a set of high-quality protein structures and evaluating how the contacts in a given protein deviate from these standard frequencies. Most often, these frequencies are translated into potentials (energies) using the Boltzmann relation (Sippl, 1990), and these 'knowledge-based' potentials are used to score the structure (for a review, see Wodak & Rooman, 1993). The potentials that consider residue–residue interactions, as in the software *PROSA II* (Sippl, 1993), are usually quite crude since each residue is represented by a single interaction centre. They can therefore detect only gross errors in chain tracing or identify incorrectly modelled segments in an otherwise correct structure, but can not validate detailed atomic positions. The same limitation applies to procedures based on three-dimensional (3D) environment profiles (Eisenberg *et al.*, 1997). The latter consider the relative frequencies of finding each of the 20 amino acids in a given local 3D environment defined by the residue buried area, the ratio of polar *versus* non-polar neighbours and the secondary structure. The corresponding energies are used to score the compatibility of a structure with its amino-acid sequence in a manner similar to the residue–residue interaction potentials.

Finally, validation procedures based on the relative frequencies of atom–atom interactions in known protein structures have also been developed (Melo & Feytmans, 1997, 1998). These methods, consolidated in the software *ANOLEA*, are capable of identifying local errors and problems of sequence misalignment in protein structures built by homology modelling. In addition, energy *Z* scores computed with these potentials for whole protein structures correlate well with the resolution of the X-ray data, as shown below for the volume-based Z scores.

21.2.2.2.3. *Deviations from standard atomic volumes as a quality measure for protein crystal structures*

The observations that protein X-ray structures are at least as tightly packed as small-molecule crystals (Richards, 1974; Harpaz *et al.*, 1994) and that the packing density inside proteins displays very limited variation (Richards, 1974; Finney, 1975) suggest that atomic volumes or measures of atomic packing can be added to the list of parameters for assessing the quality of protein structures.

Packing and related measures have been used to compare structures of proteins derived by both X-ray diffraction and NMR spectroscopy. Ratnaparkhi *et al.* (1998) analysed pairs of protein structures for which both crystal and NMR structures were available. They found that the packing values of the NMR models displayed a much larger scatter than those of the corresponding crystal structures, suggesting that this is probably due to the fact that accurate values of the packing density cannot, at present, be obtained from NMR data. Similar conclusions were reached using measures of residue–residue contact area (Abagyan & Totrov, 1997).

Here, we describe the approach of Pontius *et al.* (1996), in which deviations from standard atomic volumes are used to assess the quality of a protein model, both overall and in specific regions.

The volumes occupied by atoms and residues inside proteins can be readily computed using the Voronoi method (1908), first applied to proteins by Richards (1974) and Finney (1975). This method uses the atomic positions of the molecular model, and the volume assigned to each atom is defined as the smallest polyhedron created by the set of planes bisecting the lines joining the atom centre to those of its neighbours (Fig. 21.2.2.1).

The use of the classical Voronoi procedure is justified in the context of validation because it avoids the need to derive a consistent set of van der Waals radii for atoms in the system. Such sets are used by other volume-calculation methods in order to partition space more accurately (Richards, 1974, 1985; Gellatly & Finney, 1982). Assigning a consistent set of radii to protein atoms is, indeed, not straightforward due to the heterogeneity of the interactions within the protein (polar, ionic, non-polar) and the presence of a large variety of hetero groups.

Structure-quality assessment based on volume calculations involves computing the atomic volumes in a subset of highly resolved and refined protein structures and analysing the distributions of these volumes for different atomic types, defined according to their chemical nature and bonded environment. These distributions define the expected ranges (mean and standard deviation) for the volume of each category of atoms. Atomic volumes in a given structure are then compared to the expected ranges, and statistically significant deviations from these ranges are flagged.

The program *PROVE* (Pontius *et al.*, 1996) implements such an approach using the analytic algorithms for volume and surface-area calculations encoded in *SurVol* (Alard, 1991). It computes for each atom i in a structure its volume Z score (Z score $= |V_i^k - \overline{V^k}|/\sigma^k$), where the superscript k designates the particular atom type (*e.g.*, the C^α atom in a Leu residue), and $\overline{V^k}$ and σ^k are, respectively, the mean and standard deviation of the reference volume distribution for the corresponding atom type. These reference distributions are derived from a set of high-quality protein crystal structures using exactly the same calculation procedure (Pontius *et al.*, 1996).

Atoms with absolute Z scores > 3 are flagged as possible problem regions in the protein model, and residues containing such atoms are highlighted on graphical plots of the same type as those used by the *PROCHECK* program and on molecular models displayed using programs such as *Rasmol* (Sayle & Milner-White, 1995).

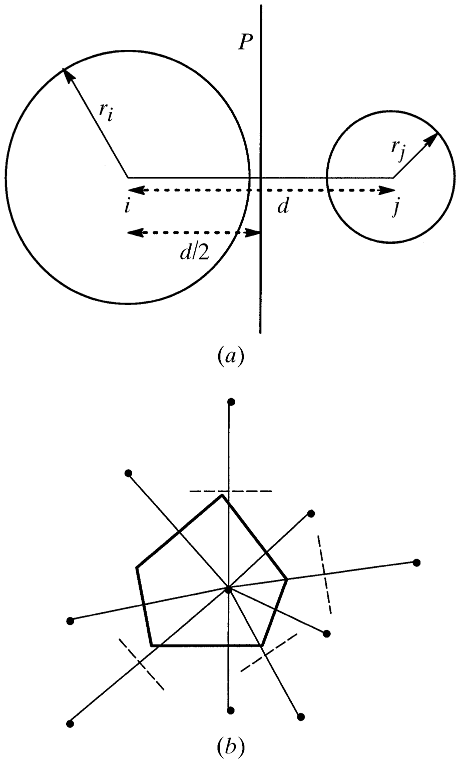

Fig. 21.2.2.1. The Voronoi polyhedron. (a) Positioning of the dividing plane P between two atoms i and j, with van der Waals radii r_i and r_j, respectively, separated by a distance d. The plane P is positioned at $d/2$. (b) 2D representation of the Voronoi polyhedron of the central atom. This polyhedron is the smallest polyhedron delimited by all the dividing planes of the atom.

In addition to the validation of the local quality of the model, its overall quality can be assessed by the *root-mean-square volume Z score* of all its atoms (see Fig. 21.2.2.2 for definition). As for many stereochemical global quality indicators, this Z score shows good correlation with the nominal resolution (d spacing) of the crystallographic data, as illustrated in Fig. 21.2.2.2(a). This figure also shows that Z-score ranges can be defined for each resolution interval. The Z scores of individual proteins that lie outside these intervals may be indicative of 'problem' structures. This is clearly the case for the two proteins 2ABX and 2GN5, whose Z scores are much higher than average (Fig. 21.2.2.2b).

Since the Voronoi volume of solvent-accessible atoms cannot be defined, because these atoms are not completely surrounded by other atoms, only completely buried atoms are scored.

The current version of *PROVE* is unable to measure the deviations from standard volumes for atoms in nucleic acids or hetero groups, simply because of the lack of reference volumes for these structures. This should change in the near future, at least for nucleic acids, thanks to the growing number of high-quality nucleic acid crystal structures from which standard volume ranges could be readily derived.

21.2.3. Validation of a model *versus* experimental data

By far the most important measure of the quality of a given atomic model is its agreement with the experimental data. This type of validation is geared towards detecting systematic errors, which determine the overall accuracy of the model, and random errors, which affect the precision of the model.

Systematic errors are difficult to detect even in highly refined structures, especially at lower resolution. The most commonly used