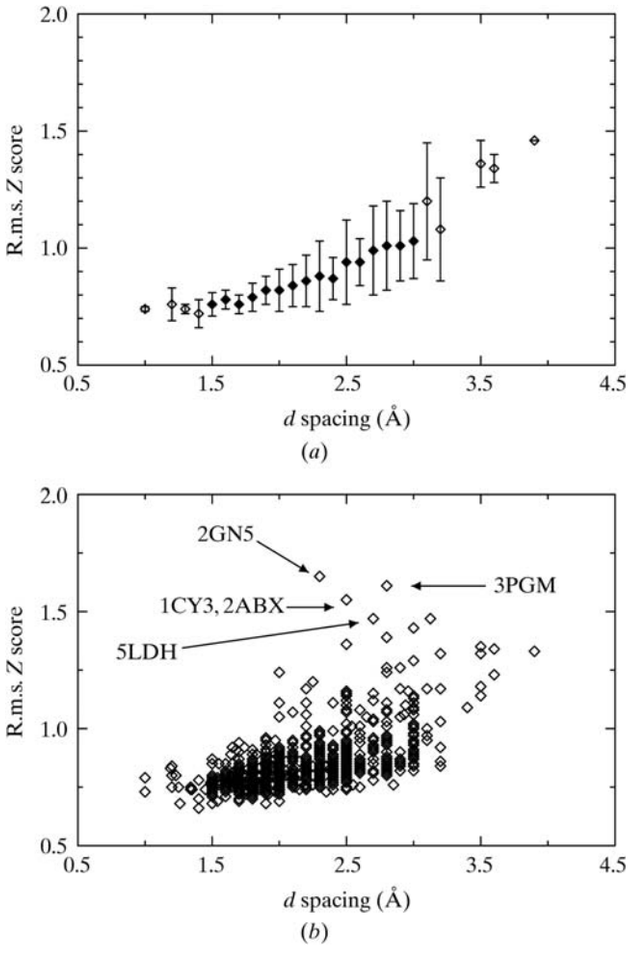

Fig. 21.2.2.2. Atomic volume Z score r.m.s. variation with nominal resolution (d spacing) in 900 protein structures from the PDB. (*a*) Average of the r.m.s. volume Z score computed for structures having the same resolution (to within ± 0.1 Å). The vertical bars indicate the magnitude of the standard deviations of the r.m.s. volume Z score in individual d-spacing bins. Graph points are derived from less than 10 structures (open diamonds) and from more than 10 structures (filled diamonds). (*b*) R.m.s. Z-score values as in (*a*), displayed for individual structures as a function of resolution. The five furthest outlier proteins are marked by their PDB codes.

measures of the agreement between the atomic coordinates and the X-ray data are the classical R factor and the 'free R factor' ($R_{\rm free}$) (Brünger, 1992*b*). The latter is based on standard statistical cross-validation techniques (Brünger, 1997) and is therefore less amenable to manipulation, such as leaving out weak data or over-fitting the data with too many parameters. Currently, nearly half of the publications on macromolecular structures report $R_{\rm free}$ values, an indication that its use is becoming more widespread. So far, however, there are no clear guidelines indicating what an 'acceptable' $R_{\rm free}$ value should be (Kleywegt & Brünger, 1996).

An expression for estimating the expected $R_{\rm free}$ value has been proposed (see Dodson *et al.*, 1996) and used to assess the significance of the drop in $R_{\rm free}$ during refinement. Accurate expressions for the expected ratio of $R_{\rm free}$ to R (the $R_{\rm free}$ ratio) have also been derived theoretically (Tickle *et al.*, 1998). This ratio seems to be independent of random errors and can be used to detect systematic errors at the convergence of the least-squares refinement. The remaining problem is to determine what the precision of $R_{\rm free}$ or the $R_{\rm free}$ ratio should be. In other words, if the $R_{\rm free}$ ratio differs from the expected value, when is the difference significant? This requires knowing the variance of these parameters. Estimating the precision of $R_{\rm free}$ can be done empirically by performing repeated refinements of the same structure with different sets of reflections removed (Brünger, 1997). From such analysis, a useful approximation to the $R_{\rm free}$ precision was suggested to be the ratio $R_{\rm free}/(n)^{1/2}$, where n is the number of reflections in the test set.

Evaluating the precision of the refined parameters, that is, the atomic coordinates and the temperature or B factors, is a different matter. In small-molecule crystallography, the standard uncertainty (s.u.) of the parameters can be computed from the variance–covariance matrix, obtained by inverting the full normal-equations matrix (Cruickshank, 1965). This can, in principle, also be done for the parameters of macromolecules. However, the number of second derivatives to be computed and the size of the matrix to be inverted are so large that this task is too time consuming to be performed routinely. This is gradually changing, however. An increasing number of proteins structures, primarily those solved at atomic resolution, have their s.u.'s computed in this manner (Deacon *et al.*, 1997; Harata *et al.*, 1998). A program often used for this purpose is *SHELXL* (Sheldrick & Schneider, 1997), a well known refinement software package for small molecules that has recently been extended to proteins. Availability of s.u.'s can determine the dependence of the precision of the atomic coordinates on various factors, such as the resolution, the atomic number, and the number and types of restraints used during refinement (Tickle *et al.*, 1998).

Other methods for determining the relative precision of atoms in macromolecular structures involve calculating the agreement between the model and the electron density in specific regions. The newer approach by Zhou *et al.* (1998) is related to the real-space R factor of Jones *et al.* (1991), but differs from it by the way in which the electron density is computed (Chapman, 1995).

As our understanding of the factors that govern the systematic errors in macromolecular crystallography increases and our ability to detect random errors improves, the possibility of devising systematic and possibly more automatic protocols for assessing the agreement between the model and the data will emerge.

In what follows, we describe the software package *SFCHECK* (Vaguine *et al.*, 1999), which can be regarded as a first attempt in this direction. This software computes and summarizes many of the commonly used measures for evaluating the quality of the structure-factor data and the agreement of the model with these data.

We summarize the tasks performed and the quality indicators computed by *SFCHECK* and briefly illustrate how this software can be used to evaluate individual structures and survey different structures.

21.2.3.1. *A systematic approach using the SFCHECK software*

21.2.3.1.1. *Tasks performed by SFCHECK*

21.2.3.1.1.1. *Treatment of structure-factor data and scaling*

SFCHECK reads in the structure-factor data written in mmCIF format. It then performs the following operations: Reflections are excluded if they are systematically absent, negative, or have flagged σ values (99.9). Equivalent reflections are merged. The amplitudes of missing reflections are approximated by taking the average value for the corresponding resolution shell.

From the model coordinates read from the PDB (or mmCIF) atomic coordinates file, *SFCHECK* calculates structure factors and scales them to the observed structure factors. The scaling factor, S, is computed using a smooth cutoff for low-resolution data (Vaguine *et al.*, 1999) (Table 21.2.3.1). This involves the calculation of the observed and calculated overall B factors from the standard deviations of the Gaussian fitted to the Patterson origin peaks [see Table 21.2.3.1 and Vaguine *et al.* (1999)]. In addition, *SFCHECK* also estimates the overall anisotropy of the data, following the

references