The general formula for the variance of the length $l = x_2 - x_1$ is

$$\sigma^2(l) = \sigma^2(x_2) - 2\mathrm{cov}(x_2, x_1) + \sigma^2(x_1). \qquad (18.5.3.13)$$

For the diffraction data alone, this gives

$$\sigma^2_{\rm diff}(l) = 1/a + 0 + 1/a = 2/a = 2\sigma^2_{\rm diff}(x_i), \qquad (18.5.3.14)$$

as expected. For the diffraction data plus restraint,

$$\begin{aligned}\sigma^2_{\rm res}(l) &= [1/a(a+2b)][(a+b) - 2b + (a+b)] \\ &= 1/(a/2 + b) \\ &< \sigma^2_{\rm diff}(l).\end{aligned} \qquad (18.5.3.15)$$

For small a, $\sigma^2_{\rm res}(l) \rightarrow 1/b = \sigma^2_{\rm geom}(l)$, as expected. The variance of the restrained length, (18.5.3.15), can be re-expressed as

$$1/\sigma^2_{\rm res}(l) = 1/\sigma^2_{\rm diff}(l) + 1/\sigma^2_{\rm geom}(l). \qquad (18.5.3.16)$$

For the two-atom protein, it can be proved directly, as one would expect from (18.5.3.16), that *restrained refinement determines a length which is the weighted mean of the diffraction-only length and the geometric dictionary length*.

The centroid has coordinate $c = (x_1 + x_2)/2$. It is easily found that $\sigma^2_{\rm res}(c) = \sigma^2_{\rm diff}(c) = 1/2a$. Thus, as expected, the restraint says nothing about the position of the molecule in the cell.

For numerical illustrations of the s.u.'s in restrained refinement, suppose the stereochemical length restraint has $\sigma_{\rm geom}(l) = 0.02$ Å. Equation (18.5.3.16) gives the length s.u. $\sigma_{\rm res}(l)$ in restrained refinement. If the diffraction-only $\sigma_{\rm diff}(x_i) = 0.01$ Å, the restrained $\sigma_{\rm res}(l)$ is 0.012 Å. If $\sigma_{\rm diff}(x_i) = 0.05$ Å, $\sigma_{\rm res}(l)$ is 0.019 Å. However large $\sigma_{\rm diff}(x_i)$, $\sigma_{\rm res}(l)$ never exceeds 0.02 Å.

Equation (18.5.3.12) gives the position s.u. $\sigma_{\rm res}(x_i)$ in restrained refinement. If the diffraction-only $\sigma_{\rm diff}(x_i) = 0.01$ Å, the restrained $\sigma_{\rm res}(x_i)$ is 0.009 Å. If $\sigma_{\rm diff}(x_i) = 0.05$ Å, $\sigma_{\rm res}(x_i) = 0.037$ Å. For large $\sigma_{\rm diff}(x_i)$, $\sigma_{\rm res}(x_i)$ tends to $\sigma_{\rm diff}(x_i)/(2)^{1/2}$ as the strong restraint couples the two atoms together. For very small $\sigma_{\rm diff}(x_i)$, the relatively weak restraint has no effect.

18.5.3.3. *Relative weighting of diffraction and restraint terms*

When only relative diffraction weights are known, as in equation (18.5.2.13), it has been common (Rollett, 1970) to scale the geometric restraint terms against the diffraction terms by replacing the restraint weights $w_{\rm geom} = 1/\sigma^2_{\rm geom}$ by $w_{\rm geom} = S^2/\sigma^2_{\rm geom}$, where $S^2 = (\sum w_h \Delta^2_h)/(n_{\rm obs} - n_{\rm params})$. However, this scheme cannot be used for low-resolution structures if $n_{\rm obs} < n_{\rm params}$.

The treatment by Tickle *et al.* (1998*a*) shows that the reduction $n_{\rm params}$ in the number of degrees of freedom has to be distributed among all the data, both diffraction observations and restraints. Since the geometric restraint weights are on an absolute scale (Å^{-2}), they propose that the (absolute) scale of the diffraction weights should be determined by adjustment until the restrained residual R' (18.5.3.1) is equal to its expected value $(n_{\rm obs} + n_{\rm restraints} - n_{\rm params})$.

For a method of determining the scale of the diffraction weights based on $R_{\rm free}$, see Brünger (1993).

The geometric restraint weights were classified by the IUCr Subcommittee (Schwarzenbach *et al.*, 1995) as derived from observations supplementary to the diffraction data, with uncertainties of type B (Section 18.5.2.3).

18.5.4. Two examples of full-matrix inversion

18.5.4.1. *Unrestrained and restrained inversions for concanavalin A*

G. M. Sheldrick extended his *SHELXL*96 program (Sheldrick & Schneider, 1997) to provide extra information about protein precision through the inversion of least-squares full matrices. His programs have been used by Deacon *et al.* (1997) for the high-resolution refinement of native concanavalin A with 237 residues, using data at 110 K to 0.94 Å refined anisotropically. After the convergence and completion of full-matrix restrained refinement for the structure, the unrestrained full matrix (coordinates only) was computed and then inverted in a massive calculation. This led to s.u's $\sigma(x)$, $\sigma(y)$, $\sigma(z)$ and $\sigma(r)$ for all atoms, and to $\sigma(l)$ and $\sigma(\theta)$ for all bond lengths and angles. $\sigma(r)$ is defined as $[\sigma^2(x) + \sigma^2(y) + \sigma^2(z)]^{1/2}$. For concanavalin A the restrained full matrix was also inverted, thus allowing the comparison of restrained and unrestrained s.u.'s.

The results for concanavalin A from the inversion of the coordinate matrices of order 6402 $(= 2134 \times 3)$ are plotted in Figs. 18.5.4.1 and 18.5.4.2. Fig. 18.5.4.1 shows $\sigma(r)$ *versus* $B_{\rm eq}$ for the fully occupied atoms of the protein (a few atoms with $B >$ 60 Å^2 are off-scale). The points are colour-coded black for carbon, blue for nitrogen and red for oxygen. Fig. 18.5.4.1(*a*) shows the restrained results, and Fig. 18.5.4.1(*b*) shows the unrestrained diffraction-data-only results. Superposed on both sets of data points are least-squares quadratic fits determined with weights $1/B^2$. At high B, the unrestrained $\sigma_{\rm diff}(r)$ can be at least double the restrained $\sigma_{\rm res}(r)$, *e.g.*, for carbon at $B = 50$ Å^2, the unrestrained $\sigma_{\rm diff}(r)$ is about 0.25 Å, whereas the restrained $\sigma_{\rm res}(r)$ is about 0.11 Å. For $B < 10$ Å^2, both $\sigma(r)$'s fall below 0.02 Å and are around 0.01 Å at $B = 6$ Å^2.

For $B < 10$ Å^2, the better precision of oxygen as compared with nitrogen, and of nitrogen as compared with carbon, can be clearly seen. At the lowest B, the unrestrained $\sigma_{\rm diff}(r)$ in Fig. 18.5.4.1(*b*) are almost as small as the restrained $\sigma_{\rm res}(r)$ in Fig. 18.5.4.1(*a*). [The quadratic fits of the restrained results in Fig. 18.5.4.1(*a*) are evidently slightly imperfect in making $\sigma_{\rm res}(r)$ tend almost to 0 as B tends to 0.]

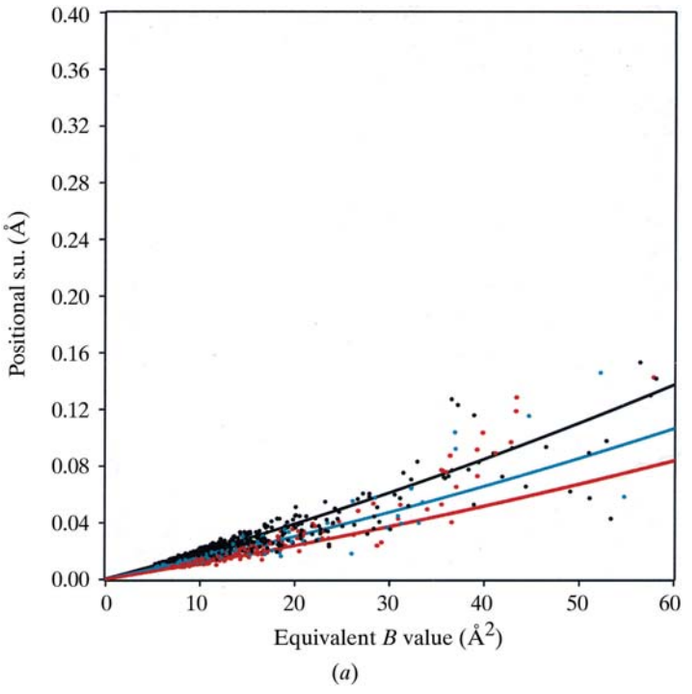
Fig. 18.5.4.2 shows $\sigma(l)$ *versus* $B_{\rm eq}$ for the bond lengths in the protein. The points are colour-coded black for C—C, blue for C—N and red for C—O. The restrained and unrestrained distributions are very different for high B. The restrained distribution in Fig. 18.5.4.2(*a*) tends to about 0.02 Å, which is the standard uncertainty of the applied restraint for 1–2 bond lengths, whereas the unrestrained distribution in Fig. 18.5.4.2(*b*) goes off the scale of the diagram. But for $B < 10$ Å^2, both distributions fall to around 0.01 Å.

The differences between the restrained and unrestrained $\sigma(r)$ and $\sigma(l)$ can be understood through the two-atom model for restrained refinement described in Section 18.5.3. For that model, the equation

$$1/\sigma^2_{\rm res}(l) = 1/\sigma^2_{\rm diff}(l) + 1/\sigma^2_{\rm geom}(l) \qquad (18.5.3.16)$$

relates the bond-length s.u. in the restrained refinement, $\sigma_{\rm res}(l)$, to the $\sigma_{\rm diff}(l)$ of the unrestrained refinement and the s.u. $\sigma_{\rm geom}(l)$ assigned to the length in the stereochemical dictionary. In the refinements, $\sigma_{\rm geom}(l)$ was 0.02 Å for all bond lengths. When this is combined in (18.5.3.16) with the unrestrained $\sigma_{\rm diff}(l)$ of any bond, the predicted restrained $\sigma_{\rm res}(l)$ is close to that found in the restrained full matrix.

It can be seen from Fig. 18.5.4.2(*b*) that many bond lengths with average $B < 10$ Å^2 have $\sigma_{\rm diff}(l) < 0.014$ Å. For these bonds the diffraction data have greater weight than the stereochemical dictionary. Some bonds have $\sigma_{\rm diff}(l)$ as low as 0.0080 Å, with



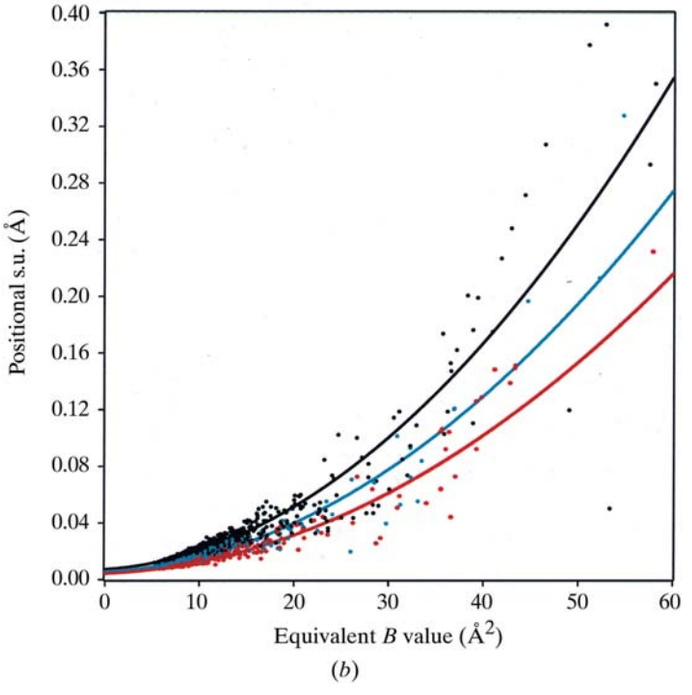
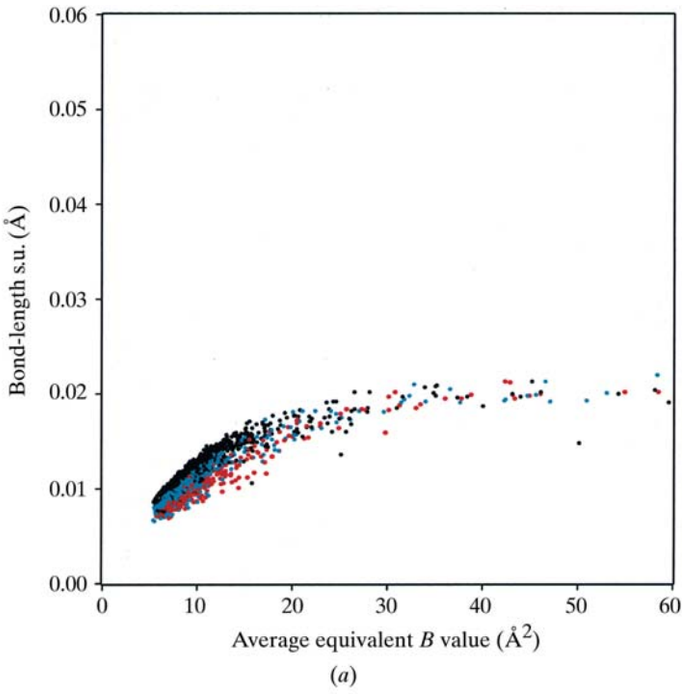


Fig. 18.5.4.1. Plots of $\sigma(r)$ *versus* B_{eq} for concanavalin A with 0.94 Å data, (*a*) restrained full-matrix $\sigma_{res}(r)$, (*b*) unrestrained full-matrix $\sigma_{diff}(r)$. Carbon black, nitrogen blue, oxygen red.



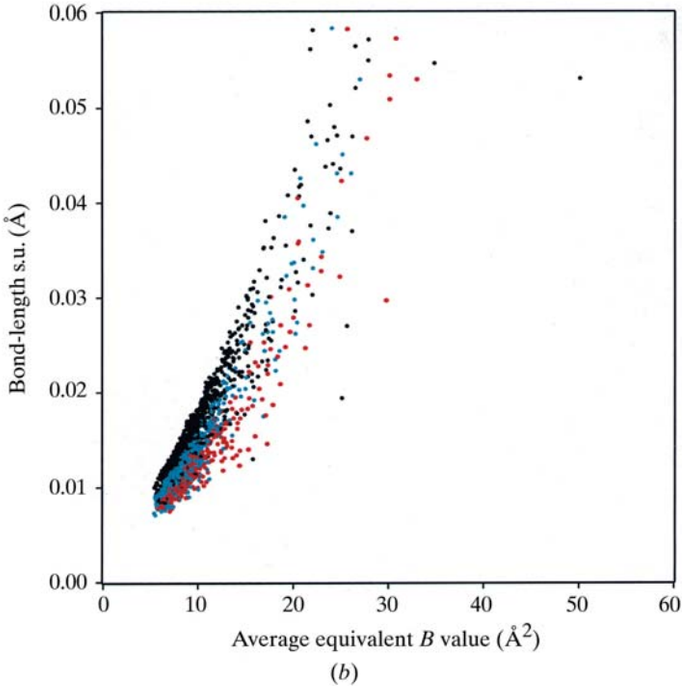


Fig. 18.5.4.2. Plots of $\sigma(l)$ *versus* average B_{eq} for concanavalin A with 0.94 Å data, (*a*) restrained full-matrix $\sigma_{res}(l)$, (*b*) unrestrained full-matrix $\sigma_{diff}(l)$. C—C black, C—N blue, C—O red.

$\sigma_{res}(l)$ around 0.0074 Å. This situation is one consequence of the availability of diffraction data to the high resolution of 0.94 Å. For large $\sigma_{diff}(l)$ (*i.e.*, high B), equation (18.5.3.16) predicts that $\sigma_{res}(l) = \sigma_{geom}(l) = 0.02$ Å, as is found in Fig. 18.5.4.2(*a*).

In an isotropic approximation, $\sigma(r) = 3^{1/2}\sigma(x)$. Equation (18.5.3.12) of the two-atom model can be recast to give

$$\sigma^2_{res}(r) = \sigma^2_{diff}(r)\left\{\left[\sigma^2_{diff}(r) + 3(0.02)^2\right] \Big/ \left[2\sigma^2_{diff}(r) + 3(0.02)^2\right]\right\}. \quad (18.5.4.1)$$

For low B, say $B \leq 15$ Å^2 in concanavalin, (18.5.4.1) gives quite good predictions of $\sigma_{res}(r)$ from $\sigma_{diff}(r)$. For instance, for a carbon atom with $B = 15$ Å^2, the quadratic curve for carbon in Fig. 18.5.4.1(*b*) shows $\sigma_{diff}(r) = 0.034$ Å, and Fig. 18.5.4.1(*a*) shows $\sigma_{res}(r) = 0.029$ Å. While if $\sigma_{diff}(r) = 0.034$ Å is used with (18.5.4.1), the resulting prediction for $\sigma_{res}(r)$ is 0.028 Å.

However, for high B, say $B = 50$ Å^2, the quadratic curve for carbon in Fig. 18.5.4.1(*b*) shows $\sigma_{diff}(r) = 0.25$ Å, and Fig. 18.5.4.1(*a*) shows $\sigma_{res}(r) = 0.11$ Å, whereas (18.5.4.1) leads to the poor estimate $\sigma_{res}(r) = 0.18$ Å.

Thus at high B, equation (18.5.4.1) from the two-atom model does not give a good description of the relationship between the

restrained and unrestrained $\sigma(r)$. The reason is obvious. Most atoms are linked by 1–2 bond restraints to two or three other atoms. Even a carbonyl oxygen atom linked to its carbon atom by a 0.02 Å restraint is also subject to 0.04 Å 1–3 restraints to chain C_α and N atoms. Consequently, for a high-B atom, when the restraints are applied it is coupled to several other atoms in a group, and its $\sigma_{\rm res}(r)$ is lower, compared with the diffraction-data-only $\sigma_{\rm diff}(r)$, by a greater amount than would be expected from the two-atom model.

18.5.4.2. *Unrestrained inversion for an immunoglobulin*

Sheldrick has provided the results of the unrestrained lower-resolution refinement of a single-chain immunoglobulin mutant (T39K) with 218 amino-acid residues, with data to 1.70 Å refined isotropically (Usón *et al.*, 1999). Fig. 18.5.4.3 shows $\sigma_{\rm diff}(r)$ *versus* $B_{\rm eq}$ for the fully occupied protein atoms. Superposed on the data points are least-squares quadratic fits. In a first very rough approximation for $\sigma_{\rm diff}(x_i)$ suggested later by equation (18.5.6.3), the dependence on atom type is controlled by $1/Z_i$, the reciprocal of the atomic number. Sheldrick found that a $1/Z_i$ dependence produced too little difference between C, N and O. The proportionalities between the quadratics for $\sigma(r)$ in Figs. 18.5.4.1 and 18.5.4.3 are based on the reciprocals of the scattering factors at $\sin\theta/\lambda = 0.3$ Å^{-1}, symbolized by $Z_i^{\#}$. For C, N and O, these are 2.494, 3.219 and 4.089, respectively. For potential use in later work, the least-squares fits to the $\sigma(r_i)Z_i^{\#}$ in Å are recorded here as

$$0.11892 + 0.00891B + 0.0001462B^2, \qquad (18.5.4.2a)$$

$$0.01826 + 0.001043B + 0.0002230B^2 \text{ and} \qquad (18.5.4.2b)$$

$$0.00115 + 0.004414B + 0.0000214B^2 \qquad (18.5.4.2c)$$

for the immunoglobulin (unrestrained), concanavalin A (unrestrained) and concanavalin A (restrained), respectively.

As might be expected from the lower resolution, the lowest $\sigma_{\rm diff}(r)$'s in the immunoglobulin are about six times the lowest $\sigma_{\rm diff}(r)$'s in concanavalin. But at $B = 50$ Å^2, the immunoglobulin curve for carbon gives $\sigma_{\rm diff}(r) = 0.37$ Å, which is only 50% larger than the concanavalin value of 0.25 Å.

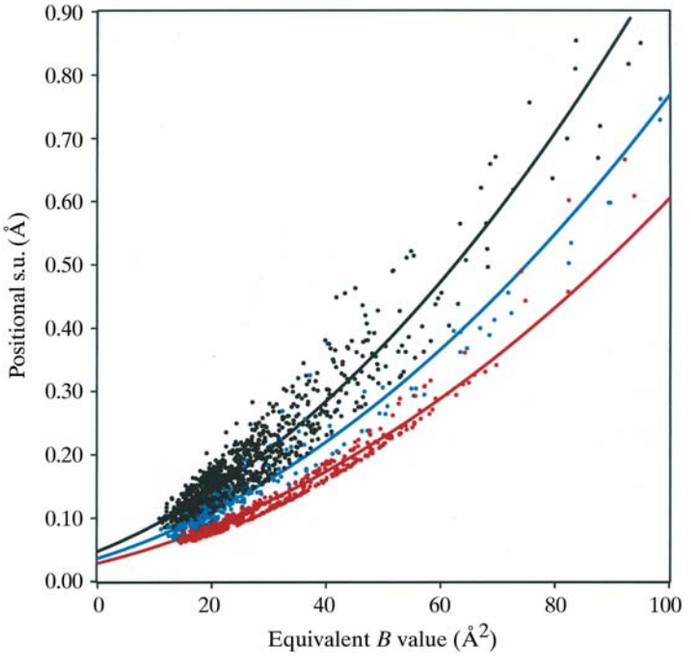


Fig. 18.5.4.3. Plot of $\sigma_{\rm diff}(r)$ *versus* $B_{\rm eq}$ from an unrestrained full matrix for immunoglobulin mutant (T39K) with 1.70 Å data. Carbon black, nitrogen blue, oxygen red.

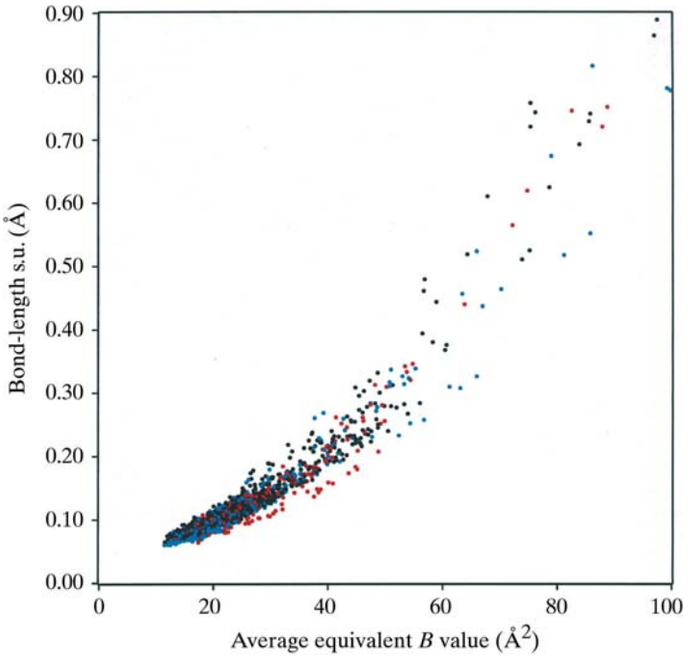


Fig. 18.5.4.4. Plot of $\sigma_{\rm diff}(l)$ *versus* average $B_{\rm eq}$ from an unrestrained full matrix for immunoglobulin mutant (T39K) with 1.70 Å data. C—O black, C—N blue, C—O red.

Fig. 18.5.4.4 shows $\sigma_{\rm diff}(l)$ *versus* $B_{\rm eq}$ for the immunoglobulin. Note that the lowest immunoglobulin unrestrained $\sigma_{\rm diff}(l)$ is about 0.06 Å, which is three times the 0.02 Å $\sigma_{\rm geom}(l)$ bond restraint.

18.5.4.3. *Comments on restrained refinement*

Geometric restraint dictionaries typically use bond-length weights based on $\sigma_{\rm geom}(l)$ of around 0.02 or 0.03 Å. Tables 18.5.7.1–18.5.7.3 show that even 1.5 Å studies have diffraction-only errors $\sigma_{\rm diff}(x, B_{\rm avg})$ of 0.08 Å and upwards. Only for resolutions of 1.0 Å or so are the diffraction-only errors comparable with the dictionary weights. Of course, the dictionary offers no values for many of the configurational parameters of the protein structure, including the centroid and molecular orientation.

18.5.4.4. *Full-matrix estimates of precision*

The opening contention of this chapter in Section 18.5.1.1 is that the variances and covariances of the structural parameters of proteins can be found from the inverse of the least-squares normal matrix. But there is a caveat, chiefly that explicit account would not be taken of disorder of the solvent or of parts of the protein. Corrections by Babinet's principle of complementarity or by mask bulk solvent models are only first-order approximations. The consequences of such disorder problems, which make the variation of calculated structure factors nonlinear over the range of interest, may in future be better handled by maximum-likelihood methods (*e.g.* Read, 1990; Bricogne, 1993*a*; Bricogne & Irwin, 1996; Murshudov *et al.*, 1997). Pannu & Read (1996) have shown how the maximum-likelihood method can be cast computationally into a form akin to least-squares calculations. Full-matrix precision estimates along the lines of the present chapter are probably somewhat low.

It should also be noted that full-matrix estimates of coordinate precision are most reliably derived from matrices involving both

coordinates and atomic displacement parameters. This is particularly important for lower-resolution analyses, in which atomic images overlap. The work on the high-resolution analysis of concanavalin A described in Section 18.5.4.1 was based on the very large coordinate matrix, of order 6402. The omission, because of computer limitations, of the anisotropic displacement parameters from the full matrix will have caused the coordinate s.u.'s of atoms with high B_{eq} to be underestimated.

Much information about the quality of a molecular model can be obtained from the eigenvalues and eigenvectors of the normal matrix (Cowtan & Ten Eyck, 2000).

18.5.5. Approximate methods

18.5.5.1. *Block calculations*

The full-matrix inversions described in the previous section require massive calculations. The length of the calculations is more a matter of the order of the matrix, *i.e.*, the number of parameters, than of the number of observations. When restraints are applied, it is the diffraction-cum-restraints full matrix which should be inverted.

With the increasing power of computers and more efficient algorithms (*e.g.* Tronrud, 1999; Murshudov *et al.*, 1999), a final full matrix should be computed and inverted much more regularly – and not just for high-resolution analyses. Low-resolution analyses have a need, beyond the indications given by B values, to identify through $\sigma(x)$ estimates their regions of tolerable and less tolerable precision.

If full-matrix calculations are impractical, partial schemes can be suggested. As far back as 1973, Watenpaugh *et al.* (1973), in a study of rubredoxin at 1.5 Å resolution, effectively inverted the diffraction full matrix in 200 parameter blocks to obtain individual s.u.'s. A similar scheme for restrained refinements could also use overlapping large blocks. A minimal block scheme in refinements of any resolution is to calculate blocks for each residue and for the block interactions between successive residues. The inversion process could then use the matrices in running groups of three successive residues, taking only the inverted elements for the central residue as the estimates of its variances and covariances.

For low-resolution analyses with very large numbers of atoms, it might be sufficient to gain a general idea of the behaviour of $\sigma(x)$ as a function of B by computing a limited number of blocks for representative or critical groups of residues. The parameters used in the blocks should include the B's, since atomic images overlap at low resolution, thus correlating the position of one atom with the displacement parameters of its neighbours.

18.5.5.2. *The modified Fourier method*

In the simplest form of the Fourier-map approach to centrosymmetric high-resolution structures, atomic positions are given by the maxima of the observed electron density. The uncertainty of such a position may be estimated as the uncertainty in the slope function (first derivative) divided by the curvature (second derivative) at the peak (Cruickshank, 1949*a*), *i.e.*,

$$\sigma(x) = \sigma(\text{slope})/(\text{atomic peak 'curvature'}). \qquad (18.5.5.1)$$

However, atomic positions are affected by finite-series and peak-overlapping effects.

Hence, more generally, atomic positions may be determined by the requirement that the slope of the difference map at the position of atom r should be zero, or equivalently that the slopes at atom r of the observed and calculated electron densities should be equal. As a criterion this becomes the basis of the modified Fourier method (Cruickshank, 1952, 1959, 1999; Bricogne, 1993*b*), which, like the least-squares method, is applicable whether or not the atomic peaks are resolved and is applicable to noncentrosymmetric structures. For refinement, a set of n simultaneous linear equations are involved, analogous to the normal equations of least squares. Their right-hand sides are the slopes of the difference map at the trial atomic positions.

The diagonal elements of the matrix, for coordinate x_r of an atom with Debye B value B_r, are approximately equal to

$$\text{'curvature'} = (4\pi^2/a^2V)\left[\sum_{hkl}(m/2)h^2 f_r \exp(-B_r \sin^2\theta/\lambda^2)\right], \qquad (18.5.5.2)$$

where $m = 1$ or 2 for acentric or centric reflections. The summation is over all independent planes and their symmetry equivalents. Strictly speaking, (18.5.5.2) is a curvature only for centrosymmetric structures.

In the modified Fourier method,

$$\sigma(\text{slope}) = (2\pi/aV)\left[\sum_{hkl} h^2(\Delta|F|^2)\right]^{1/2}. \qquad (18.5.5.3)$$

This is simply an estimate of the r.m.s. uncertainty at a general position (Cruickshank & Rollett, 1953) in the slope of the difference map, *i.e.*, the r.m.s. uncertainty on the right-hand side of the modified Fourier method.

$\sigma(x)$ is then given by (18.5.5.1), using (18.5.5.3) and (18.5.5.2).

18.5.5.3. *Application of the modified Fourier method*

An extreme example of an apparently successful gross approximation to protein precision is represented by Daopin *et al.*'s (1994) treatment of two independent determinations (at 1.8 and 1.95 Å) of the structure of TGF-β2. They reported that the modified Fourier-map formulae given in Section 18.5.5.2 yielded a quite good description of the B dependence of the positional differences between the two independent determinations. However, there is a formal difficulty about this application. Equation (18.5.5.1) derives from a diffraction-data-only approach, whereas the two structures were determined from restrained refinements. Even though the *TNT* restraint parameters and weights may have been the same in both refinements, it is slightly surprising that (18.5.5.1) should have worked well.

Equation (18.5.2.1) requires the summation of various series over all (hkl) observations; such calculations are not customarily provided in protein programs. However, due to the fundamental similarities between Fourier and least-squares methods demonstrated by Cochran (1948), Cruickshank (1949*b*, 1952, 1959), and Cruickshank & Robertson (1953), closely similar estimates of the precision of individual atoms can be obtained from the reciprocal of the diagonal elements of the diffraction-data-only least-squares matrix. These elements will often have been calculated already within the protein refinement programs, but possibly never output. Such estimates could be routinely available.

Between approximations using largish blocks and those using only the reciprocals of diagonal terms, a whole variety of intermediate approximations involving some off-diagonal terms could be envisaged.

Whatever method is used to estimate uncertainties, it is essential to distinguish between *coordinate* uncertainty, *e.g.*, $\sigma(x)$, and *position* uncertainty $\sigma(r) = [\sigma^2(x) + \sigma^2(y) + \sigma^2(z)]^{1/2}$.

The remainder of this chapter discusses two rough-and-ready indicators of structure precision: the diffraction-component precision index (DPI) and Luzzati plots.

references