

18. REFINEMENT

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

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

For the diffraction data alone, this gives

$$\sigma_{\text{diff}}^2(l) = 1/a + 0 + 1/a = 2/a = 2\sigma_{\text{diff}}^2(x_i), \quad (18.5.3.14)$$

as expected. For the diffraction data plus restraint,

$$\begin{aligned} \sigma_{\text{res}}^2(l) &= [1/a(a + 2b)][(a + b) - 2b + (a + b)] \\ &= 1/(a/2 + b) \\ &< \sigma_{\text{diff}}^2(l). \end{aligned} \quad (18.5.3.15)$$

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

$$1/\sigma_{\text{res}}^2(l) = 1/\sigma_{\text{diff}}^2(l) + 1/\sigma_{\text{geom}}^2(l). \quad (18.5.3.16)$$

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 $\sigma_{\text{res}}^2(c) = \sigma_{\text{diff}}^2(c) = 1/2a$. 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 $\sigma_{\text{geom}}(l) = 0.02 \text{ \AA}$. Equation (18.5.3.16) gives the length s.u. $\sigma_{\text{res}}(l)$ in restrained refinement. If the diffraction-only $\sigma_{\text{diff}}(x_i) = 0.01 \text{ \AA}$, the restrained $\sigma_{\text{res}}(l)$ is 0.012 \AA . If $\sigma_{\text{diff}}(x_i) = 0.05 \text{ \AA}$, $\sigma_{\text{res}}(l)$ is 0.019 \AA . However large $\sigma_{\text{diff}}(x_i)$, $\sigma_{\text{res}}(l)$ never exceeds 0.02 \AA .

Equation (18.5.3.12) gives the position s.u. $\sigma_{\text{res}}(x_i)$ in restrained refinement. If the diffraction-only $\sigma_{\text{diff}}(x_i) = 0.01 \text{ \AA}$, the restrained $\sigma_{\text{res}}(x_i)$ is 0.009 \AA . If $\sigma_{\text{diff}}(x_i) = 0.05 \text{ \AA}$, $\sigma_{\text{res}}(x_i) = 0.037 \text{ \AA}$. For large $\sigma_{\text{diff}}(x_i)$, $\sigma_{\text{res}}(x_i)$ tends to $\sigma_{\text{diff}}(x_i)/(2)^{1/2}$ as the strong restraint couples the two atoms together. For very small $\sigma_{\text{diff}}(x_i)$, 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 $w_{\text{geom}} = 1/\sigma_{\text{geom}}^2$ by $w_{\text{geom}} = S^2/\sigma_{\text{geom}}^2$, where $S^2 = (\sum w_h \Delta_h^2)/(n_{\text{obs}} - n_{\text{params}})$. However, this scheme cannot be used for low-resolution structures if $n_{\text{obs}} < n_{\text{params}}$.

The treatment by Tickle *et al.* (1998a) shows that the reduction n_{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 (\AA^{-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 $(n_{\text{obs}} + n_{\text{restraints}} - n_{\text{params}})$.

For a method of determining the scale of the diffraction weights based on R_{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 *SHELXL96* 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 \AA 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_{eq} for the fully occupied atoms of the protein (a few atoms with $B > 60 \text{ \AA}^2$ 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 $\sigma_{\text{diff}}(r)$ can be at least double the restrained $\sigma_{\text{res}}(r)$, e.g., for carbon at $B = 50 \text{ \AA}^2$, the unrestrained $\sigma_{\text{diff}}(r)$ is about 0.25 \AA , whereas the restrained $\sigma_{\text{res}}(r)$ is about 0.11 \AA . For $B < 10 \text{ \AA}^2$, both $\sigma(r)$'s fall below 0.02 \AA and are around 0.01 \AA at $B = 6 \text{ \AA}^2$.

For $B < 10 \text{ \AA}^2$, 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 $\sigma_{\text{diff}}(r)$ in Fig. 18.5.4.1(b) are almost as small as the restrained $\sigma_{\text{res}}(r)$ 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 $\sigma_{\text{res}}(r)$ tend almost to 0 as B tends to 0.]

Fig. 18.5.4.2 shows $\sigma(l)$ versus B_{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 \AA , 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 $B < 10 \text{ \AA}^2$, both distributions fall to around 0.01 \AA .

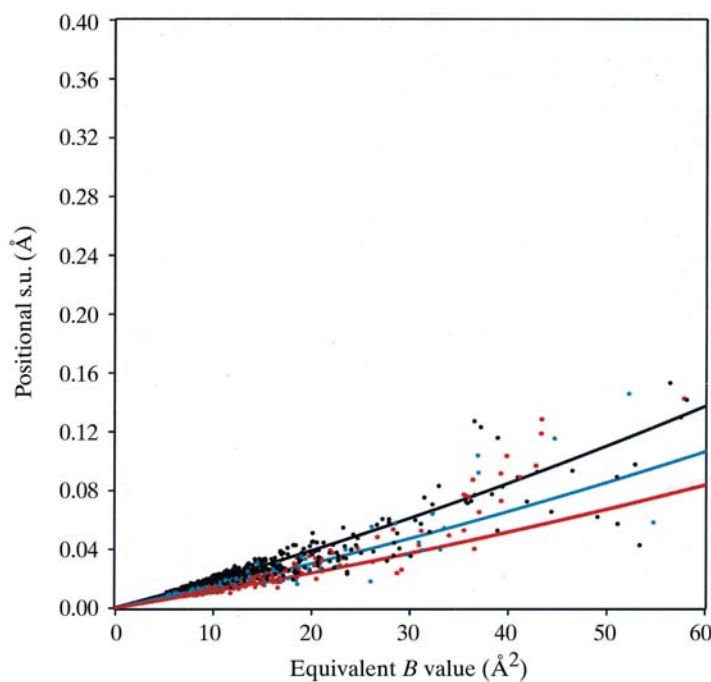
The differences between the restrained and unrestrained $\sigma(r)$ and $\sigma(l)$ can be understood through the two-atom model for restrained refinement described in Section 18.5.3. For that model, the equation

$$1/\sigma_{\text{res}}^2(l) = 1/\sigma_{\text{diff}}^2(l) + 1/\sigma_{\text{geom}}^2(l) \quad (18.5.3.16)$$

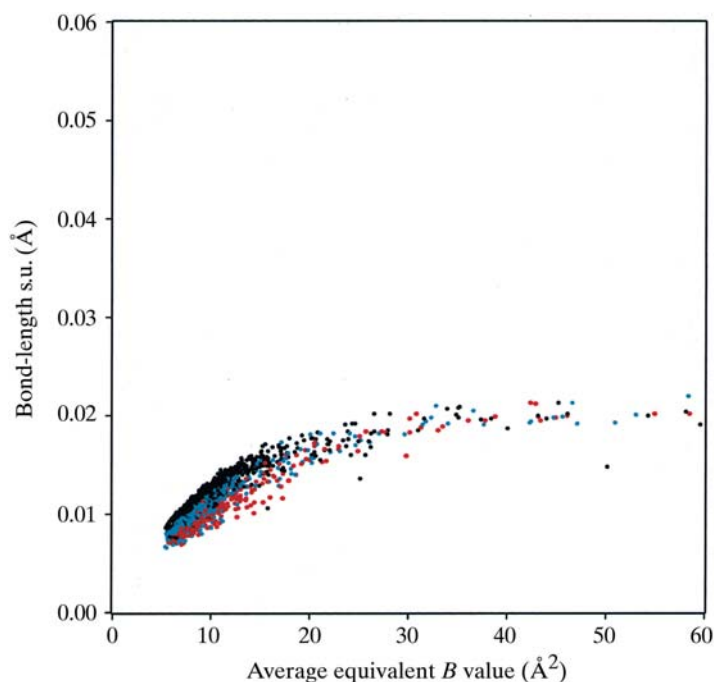
relates the bond-length s.u. in the restrained refinement, $\sigma_{\text{res}}(l)$, to the $\sigma_{\text{diff}}(l)$ of the unrestrained refinement and the s.u. $\sigma_{\text{geom}}(l)$ assigned to the length in the stereochemical dictionary. In the refinements, $\sigma_{\text{geom}}(l)$ was 0.02 \AA for all bond lengths. When this is combined in (18.5.3.16) with the unrestrained $\sigma_{\text{diff}}(l)$ of any bond, the predicted restrained $\sigma_{\text{res}}(l)$ 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 $B < 10 \text{ \AA}^2$ have $\sigma_{\text{diff}}(l) < 0.014 \text{ \AA}$. For these bonds the diffraction data have greater weight than the stereochemical dictionary. Some bonds have $\sigma_{\text{diff}}(l)$ as low as 0.0080 \AA , with

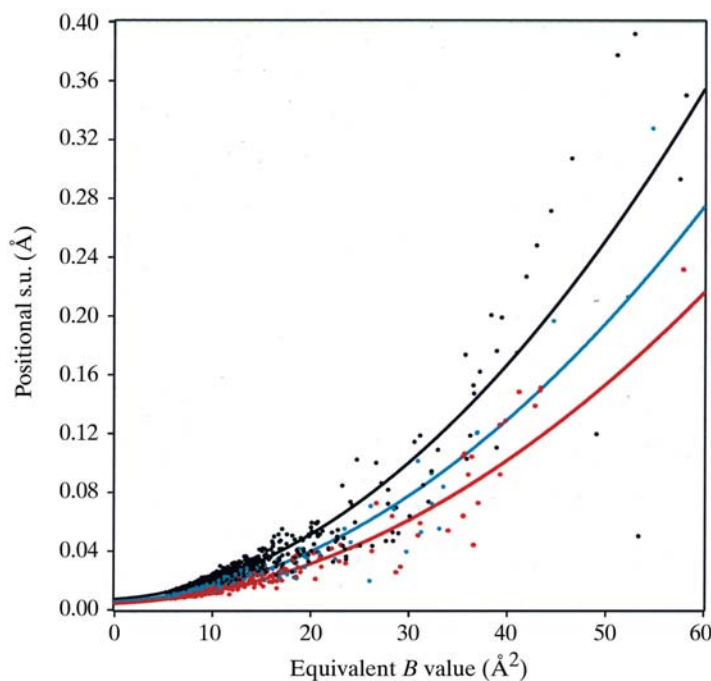
18.5. COORDINATE UNCERTAINTY



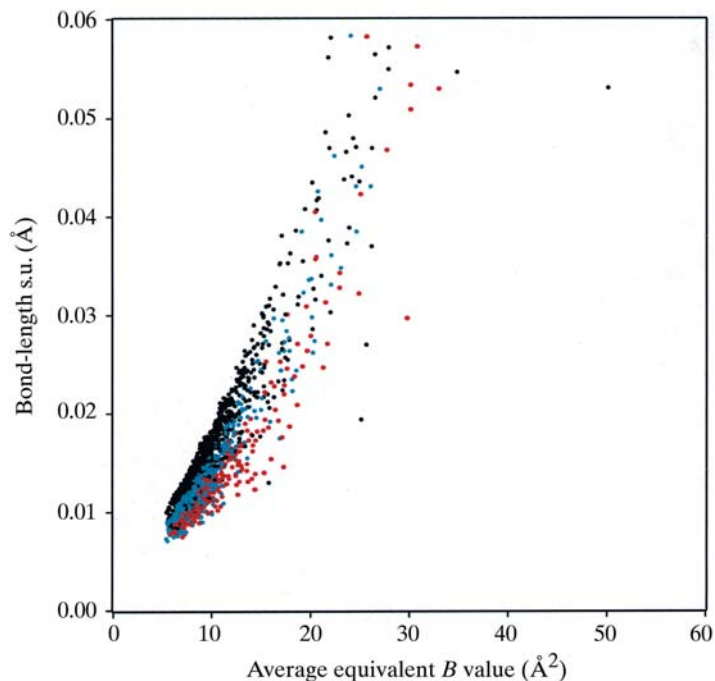
(a)



(a)



(b)



(b)

Fig. 18.5.4.1. Plots of $\sigma(r)$ versus B_{eq} for concanavalin A with 0.94 Å data, (a) restrained full-matrix $\sigma_{\text{res}}(r)$, (b) unrestrained full-matrix $\sigma_{\text{diff}}(r)$. 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 $\sigma_{\text{res}}(l)$, (b) unrestrained full-matrix $\sigma_{\text{diff}}(l)$. C—C black, C—N blue, C—O red.

$\sigma_{\text{res}}(l)$ 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 $\sigma_{\text{diff}}(l)$ (i.e., high B), equation (18.5.3.16) predicts that $\sigma_{\text{res}}(l) = \sigma_{\text{geom}}(l) = 0.02$ 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

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

For low B , say $B \leq 15 \text{ Å}^2$ in concanavalin, (18.5.4.1) gives quite

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

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

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 $\sigma_{\text{res}}(r)$ is lower, compared with the diffraction-data-only $\sigma_{\text{diff}}(r)$, 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 $\sigma_{\text{diff}}(r)$ versus B_{eq} for the fully occupied protein atoms. Superposed on the data points are least-squares quadratic fits. In a first very rough approximation for $\sigma_{\text{diff}}(x_i)$ 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 $\sin \theta/\lambda = 0.3 \text{ \AA}^{-1}$, 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

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

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

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

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

As might be expected from the lower resolution, the lowest $\sigma_{\text{diff}}(r)$'s in the immunoglobulin are about six times the lowest $\sigma_{\text{diff}}(r)$'s in concanavalin. But at $B = 50 \text{ \AA}^2$, the immunoglobulin

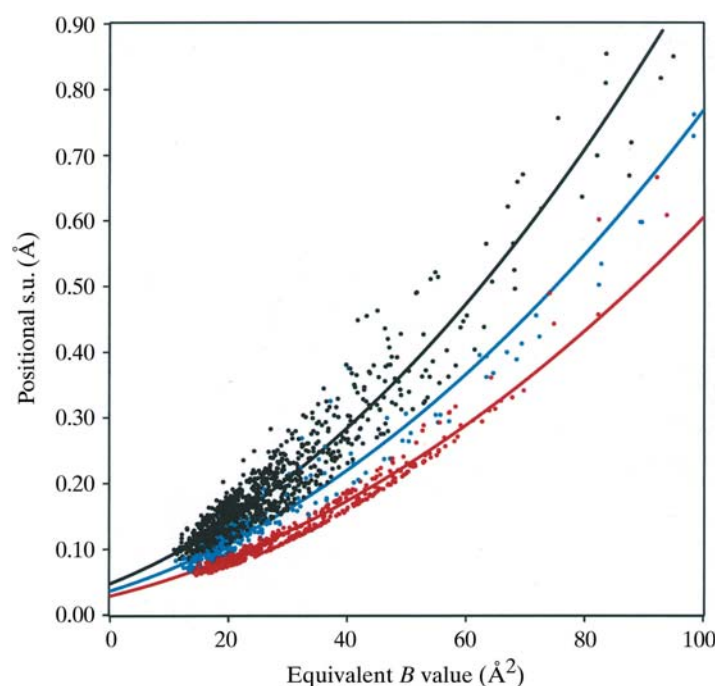


Fig. 18.5.4.3. Plot of $\sigma_{\text{diff}}(r)$ versus B_{eq} from an unrestrained full matrix for immunoglobulin mutant (T39K) with 1.70 Å data. Carbon black, nitrogen blue, oxygen red.

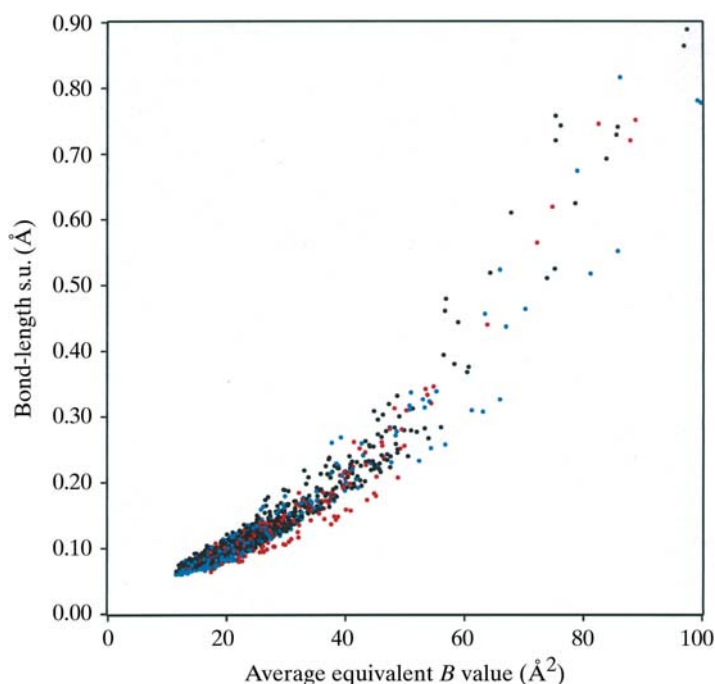


Fig. 18.5.4.4. Plot of $\sigma_{\text{diff}}(l)$ versus average B_{eq} from an unrestrained full matrix for immunoglobulin mutant (T39K) with 1.70 Å data. C—O black, C—N blue, C—O red.

curve for carbon gives $\sigma_{\text{diff}}(r) = 0.37 \text{ \AA}$, which is only 50% larger than the concanavalin value of 0.25 Å.

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

18.5.4.3. Comments on restrained refinement

Geometric restraint dictionaries typically use bond-length weights based on $\sigma_{\text{geom}}(l)$ 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 $\sigma_{\text{diff}}(x, B_{\text{avg}})$ 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, 1993a; 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