Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Remarks about protein structure precision

D. W. J. Cruickshank†

Chemistry Department, UMIST, Manchester M60 1QD, England

† Address for correspondence: D. W. J. Cruickshank, 105 Moss Lane, Alderley Edge, Cheshire SK9 7HW, England.

Correspondence e-mail: dwj_cruickshank@email.msn.com

Durward Cruickshank joined E. G. Cox's chemical crystallography group at Leeds University in 1946. Cox was a pioneer from 1937 onwards of X-ray structure analysis with three-dimensional data, and the Leeds laboratory was strong in computing by punched-card methods. Cruickshank published on the accuracy of crystal structure analysis by Fourier and least-squares methods from 1949 onwards. In September 1950 Cox sent him to the world's first summer school on electronic computing run at the EDSAC laboratory in Cambridge. From 1952 to 1957 the Leeds group made very heavy use of the Ferranti Mark I computer at Manchester University. This machine was the commercial version of the first electronic stored-program computer, built and run by F. C. Williams and T. Kilburn in 1948. The computation of Booth's differential syntheses for structure refinement led directly to Cruickshank's 1956 papers in *Acta Crystallographica* on anisotropic atomic vibrations, molecular rigid-body vibrations and libration corrections and to the refinement of J. M. Robertson's structure for anthracene. His later interests have included the structures of second-row oxides, gas-phase electron diffraction and theoretical chemistry. From 1962 to 1967 he was the colleague of Robertson at Glasgow University as Joseph Black Professor of Chemistry. Since 1967 he has been at the University of Manchester Institute of Science and Technology. Early retirement in 1983 has allowed him time for research as a hobby. He had a happy collaboration with John Helliwell and Keith Moffat in the revival of the Laue method as a tool in time-resolved macromolecular crystallography. With N. Kato and H. J. Juretschke he edited the 1992 IUCr Memorial Volume for P. P. Ewald. He has been IUCr Treasurer (1966–1972) and General Secretary (1970–1972).

Received 17 April 1998
Accepted 30 September 1998

Full-matrix least squares is taken as the basis for an examination of protein structure precision. A two-atom protein model is used to compare the precisions of unrestrained and restrained refinements. In this model, restrained refinement determines a bond length which is the weighted mean of the unrestrained diffraction-only length and the geometric dictionary length. Data of 0.94 Å resolution for the 237-residue protein concanavalin A are used in unrestrained and restrained full-matrix inversions to provide standard uncertainties $\sigma(r)$ for positions and $\sigma(l)$ for bond lengths. $\sigma(r)$ is as small as 0.01 Å for atoms with low Debye B values but increases strongly with B. The results emphasize the distinction between unrestrained and restrained refinements and between $\sigma(r)$ and $\sigma(l)$. Other full-matrix inversions are reported. Such inversions require massive calculations. Several approximate methods are examined and compared critically. These include a Fourier map formula [Cruickshank (1949). *Acta Cryst.* **2**, 65–82], Luzzati plots [Luzzati (1952). *Acta Cryst.* **5**, 802–810] and a new diffraction-component precision index (DPI). The DPI estimate of $\sigma(r, B_{\rm avg})$ is given by a simple formula. It uses R or $R_{\rm free}$ and is based on a very rough approximation to the least-squares method. Many examples show its usefulness as a precision comparator for high- and low-resolution structures. The effect of restraints as resolution varies is examined. More regular use of full-matrix inversion is urged to establish positional precision and hence the precision of non-dictionary distances in both high- and low-resolution structures. Failing this, parameter blocks for representative residues and their neighbours should be inverted to gain a general idea of $\sigma(r)$ as a function of B. The whole discussion is subject to some caveats about the effects of disordered regions in the crystal.

1. Introduction

1.1. Background

These remarks were prompted by the numerous papers on protein structures which report the estimation of final errors by Luzzati (1952) plots of R *versus* $2\sin\theta/\lambda$. Unfortunately, Luzzati developed his elegant theory for a quite different purpose, and the use of Luzzati plots to estimate final errors in protein structures is often badly flawed. A critical discussion of Luzzati's theory will be offered in §§8, 9 and 10. However, plots of R *versus* $2\sin\theta/\lambda$ remain valuable.

Just over 50 years ago, E. G. Cox and G. A. Jeffrey started my interest in the accuracy of the structures of small molecules as determined by X-ray crystallography (Cox & Cruickshank, 1948; Cruickshank, 1949*a*). Recently, I became interested in protein accuracy, not only because of the misuse of Luzzati plots, but also because Daopin *et al.* (1994), in a paper on the

accuracy of two structures of TGF-β2, made generous remarks about error formulae of mine dating back to 1949.

Even in 1967, when the first few protein structures had been solved, it would have been hard to imagine that a time would come when the best protein structures would be determined with a precision approaching that of small molecules. That time was reached some while ago. Consequently, the methods for the assessment of the precision of small molecules can be extended to good-quality protein structures.

The key idea is simply stated. At the conclusion and full convergence of a least-squares or equivalent refinement, *the estimated variances and covariances of the parameters may be obtained through the inversion of the least-squares full matrix*.

The inversion of the full matrix for a large protein is a gigantic computational task, but it is being accomplished in an increasing number of cases. Alternatively, approximations may be sought. Often these can be no more than rough order-of-magnitude estimates. Some of these approximations are considered below.

Caveat. Quite apart from their large numbers of atoms, protein structures show features differing from well ordered small-molecule structures. Protein crystals contain large amounts of solvent, much of it not well ordered. Parts of the protein chain may be floppy or disordered. All natural protein crystals are non-centrosymmetric, hence the simplifications of error assessment for centrosymmetric structures are inapplicable. The effects of incomplete modelling of disorder on phase angles, and thus on parameter errors, are not addressed explicitly in the following analysis. Nor does this analysis address the quite different problem of possible gross errors or misplacements in a structure, other than by their indication through high B values or high coordinate standard uncertainties (s.u.s, formerly called estimated standard deviations).

Some of the structure determinations reported in this paper do make a first-order correction for the effects of disordered solvent on phase angles by application of Babinet's principle of complementarity (Langridge *et al.*, 1960). Babinet's principle follows from the fact that if $\rho(\mathbf{x})$ is constant throughout the cell, then $F(\mathbf{h}) = 0$, except for $F(0)$. Consequently, if the cell is divided into two regions C and D, $F_C(\mathbf{h}) = -F_D(\mathbf{h})$. Thus, if D is a region of disordered solvent, $F_D(\mathbf{h})$ can be estimated from $-F_C(\mathbf{h})$. A first approximation to a disordered model may be obtained by placing negative point-atoms with very high Debye B values at all the ordered sites in region C. This procedure provides some correction for very low resolution planes.

The application of restraints in protein refinement does not affect the key idea about the method of error estimation. A simple model for restrained refinement is analysed in §3, and the effect of restraints is discussed in §4 and later.

This paper is not offered as a comprehensive discussion of protein precision, but as a pointer to some useful possibilities and as a stimulus to further work by others. Preliminary accounts of some of the material appeared in the Report of a Workshop held in York in 1995 (Dodson *et al.*, 1996), in the Report of a CCP4 Study Weekend (Cruickshank, 1996*a*) and in the abstract for a poster at the IUCr Congress in Seattle (Cruickshank, 1996*b*).

Protein structures which exhibit non-crystallographic symmetry are not considered in this paper.

1.2. Accuracy and precision

A distinction should be made between the terms *accuracy* and *precision*. A single measurement of the magnitude of a quantity differs by error from its unknown true value λ. In statistical theory (Cruickshank, 1959), the fundamental supposition made about errors is that for a given experimental procedure, the possible results of an experiment define the probability density function $f(x)$ of a *random variable*. Both the true value λ and the probability density $f(x)$ are unknown. The problem of assessing the accuracy of a measurement is thus the double problem of estimating $f(x)$ and of assuming a relation between $f(x)$ and λ.

Precision relates to the function $f(x)$ and its spread.

The problem of what relation to assume between $f(x)$ and the true value λ is more subtle, involving particularly the question of *systematic errors*. The usual procedure, after correcting for known systematic errors, is to suppose that some typical property of $f(x)$, often the mean, is the value of λ. No repetition of the same experiment will ever reveal the systematic errors, so that statistical estimates of precision take into account only random errors. Empirically, systematic errors can be detected only by remeasuring the quantity with a different technique.

In older papers, the word accuracy is often intended to cover both random and systematic errors or it may cover only random errors in the sense of precision (known systematic errors having been corrected). In this paper, except when summarizing older work, I have generally avoided accuracy when precision is meant.

As much of the basic material about precision estimates by the least-squares and Fourier methods is dispersed in the older literature, an outline summary of some key features is given in the *Appendix*.

2. Effect of atomic displacement parameters (or 'temperature factors')

It is useful to begin with a reminder that the Debye $B = 8\pi^2\langle u^2\rangle$, where u is the atomic displacement parameter. If B = 80 Å^2, the r.m.s. amplitude is 1.01 Å. The centroid of an atom with such a B is unlikely to be precisely determined. For B = 40 Å^2, the r.m.s. amplitude of an atom, 0.71 Å, is approximately half a C—N bond length. For B = 20 Å^2, the amplitude is 0.50 Å. Even for B = 5 Å^2, the amplitude is 0.25 Å. The size of the atomic displacement amplitudes should always be borne in mind when considering the precision of the position of the centroid of an atom.

Scattering power depends on $\exp[-2B(\sin\theta/\lambda)^2] = \exp[-B/(2d^2)]$. For B = 20 Å^2 and d = 4, 2 or 1 Å, this factor is 0.54, 0.08 or 0.0001, respectively. For d = 2 Å and B = 5, 20 or 80 Å^2, the factor is again 0.54, 0.08 or 0.0001, respectively. The

scattering power of an atom thus depends very strongly on B and on the resolution $d = 1/s = \lambda/2\sin\theta$. Scattering at high resolution (low d) is dominated by atoms with low B.

[An IUCr Subcommittee (Trueblood *et al.*, 1996) has recently recommended that the phrase 'temperature factor', though widely used in the past, should be avoided on account of several ambiguities in its meaning and usage. The Subcommittee also discourages the use of B and the anisotropic tensor **B** in favour of $\langle u^2\rangle$ and **U**, on the grounds that the latter have a more direct physical significance. The present author concurs (Cruickshank, 1956, 1965). However, as the use of B or B_{eq} is currently so widespread in biomolecular crystallography, this paper has been written in terms of B.]

Important papers on the accuracy of refined protein structures have been published by Chambers & Stroud (1979) and Daopin *et al.* (1994). Chambers & Stroud compared two independently refined models of bovine trypsin, while Daopin *et al.* compared two structures of TGF-β2. A number of rather similar points were made in both papers. For simplicity, only the more recent paper by Daopin *et al.* will be summarized here.

The structure of transforming growth factor β2 with 112 amino acids was determined independently by Daopin & Davies and by Schlunegger & Grütter. The sources of the two samples of TGF-β2, named 1TGI and 1TGF, were very different. Both have space group $P3_221$ with nearly identical unit-cell dimensions. Different heavy-atom derivatives were used to provide the initial phases. The refinements of x, y, z and isotropic B for the non-H atoms were performed with the same program package *TNT* at comparable resolutions of 1.8 and 1.95 Å, respectively. Final residual R factors were 0.173 and 0.188, respectively. Protein atoms totalled 890 in both refinements, with 58 and 84 water molecules, respectively.

Structural comparison showed that the two structures were nearly identical, with the differences mostly in the mobile region. The r.m.s. differences in position between the two structures were 0.10 Å for 104 pairs of Cα atoms, 0.15 Å for 434 pairs of main-chain atoms and 0.33 Å for 860 out of 890 pairs of protein atoms.

The authors plotted the r.m.s. position differences $\langle\Delta r\rangle$ between the Cα atoms in the two structures *versus* residue number, and showed that these structural differences were highly correlated with the Debye B factors. This provided another direct demonstration that atomic precision in proteins depends strongly on B. They then showed (Fig. 1) that the agreement between the r.m.s. structure differences $\langle\Delta r\rangle$ and the errors $\sigma(r)$, *i.e.* standard uncertainties, estimated by a formula of Cruickshank (1949*a*, 1952, 1959) was 'quite good throughout the entire range of B values'.

This formula, based on a Fourier map approach, can be described approximately as

$$\sigma(x) = \sigma(\text{slope})/(\text{atomic peak 'curvature'}). \qquad (1)$$

The σ(slope) term is the same for all atoms and is proportional to

$$\left(\sum_{\text{obs}} h^2|\Delta F|^2\right)^{1/2}. \qquad (2)$$

The second derivative 'curvature' term, which depends on B and the atom type i, is proportional to

$$\sum_{\text{obs}} h^2 f_i(\sin\theta/\lambda)[\exp(-B\sin^2\theta/\lambda^2)](m/2), \qquad (3)$$

where m = 1 or 2 for acentric or centric reflections. Thus, $\sigma(x)$ increases steadily with B, as found.

Chambers & Stroud (1979) considered that the Cruickshank (1949*a*) formula, while giving a strong B dependence, underestimated the real errors.

An indication of the derivation of (1) is given in the *Appendix*, §A.2. Equation (1) is not a least-squares formula, but it is closely related.

3. Restrained refinement

3.1. Residual function

Proteins are usually refined by a restrained refinement program such as *PROLSQ* (Hendrickson & Konnert, 1980). Here, a function of the type

$$R' = \sum w_h(\Delta F)^2 + \sum w_{\text{geom}}(\Delta Q)^2 \qquad (4)$$

is minimized, where Q denotes a geometrical restraint such as a bond length. Formally, all one is doing is extending the list of observations. One is adding to the protein diffraction data geometrical data from a stereochemical dictionary such as that of Engh & Huber (1991). A chain C—N bond length may be known from the dictionary with much greater precision $1/w_{\text{geom}}^{1/2}$, say 0.02 Å, than from an unrestrained diffraction-data-only protein refinement.

In a high-resolution unrestrained refinement of a small molecule, the standard uncertainty (s.u.) of a bond length $A—B$ is often well approximated by

$$\sigma(l) = (\sigma_A^2 + \sigma_B^2)^{1/2}. \qquad (5)$$

However, in a protein determination $\sigma(l)$ is often much smaller than either σ_A or σ_B because of the excellent information from the stereochemical dictionary which correlates the positions of A and B.

Laying aside computational size and complexity, the protein precision problem is straightforward in principle. When a restrained refinement has converged to an acceptable structure and the shifts in successive rounds have become negligible, invert the full matrix. The inverse matrix immediately yields estimates of the variances and covariances of all parameters.

(The dimensions of the matrix are the same whether the refinement is restrained or not. The full matrix will be rather sparse, but not nearly as sparse as in a small-molecule refinement. For present purposes, it is irrelevant whether the residual for the diffraction data is based on $|F|$ or $|F|^2$. For comment on the relative weighting of the diffraction and restraint terms, see the *Appendix* §A.3.)

3.2. A very simple protein model

Some aspects of restrained refinement are easily understood by considering a *one-dimensional protein consisting of two like atoms in the asymmetric unit*, with coordinates x_1 and x_2 relative to a fixed origin and bond length $l = x_2 - x_1$. In the refinement, the normal equations are of the type $\mathbf{N}\Delta\mathbf{x} = \mathbf{e}$. For two non-overlapping like atoms, the *diffraction* data will yield a normal matrix

$$\mathbf{N} = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}, \tag{6}$$

with inverse

$$\begin{pmatrix} 1/a & 0 \\ 0 & 1/a \end{pmatrix}, \tag{7}$$

where

$$a = \sum w_h(\partial F_h/\partial x_i)^2. \tag{8}$$

A *geometric restraint* on the length will yield a normal matrix

$$\begin{pmatrix} b & -b \\ -b & b \end{pmatrix}, \tag{9}$$

with no inverse since its determinant is zero, where

$$b = w_{\rm geom}(\partial l/\partial x_i)^2. \tag{10}$$

Note $\partial l/\partial x_2 = -\partial l/\partial x_1 = 1$, so that

$$b = w_{\rm geom} = 1/\sigma^2_{\rm geom}(l), \tag{11}$$

where $\sigma^2_{\rm geom}(l)$ is the variance assigned to the length in the stereochemical dictionary.

Combining the diffraction data and the restraint, the normal matrix becomes

$$\begin{pmatrix} a+b & -b \\ -b & a+b \end{pmatrix}, \tag{12}$$

with inverse

$$\{1/[a(a+2b)]\}\begin{pmatrix} a+b & b \\ b & a+b \end{pmatrix}. \tag{13}$$

For the diffraction data alone, the variance of x_i is

$$\sigma^2_{\rm diff}(x_i) = 1/a. \tag{14}$$

For the diffraction data plus restraint, the variance of x_i is

$$\sigma^2_{\rm res}(x_i) = (a+b)/[a(a+2b)] < \sigma^2_{\rm diff}(x_i). \tag{15}$$

Note that though the restraint says nothing about the position of x_i, the variance of x_i has been reduced because of the coupling to the position of the other atom. In the limit when $a << b$, $\sigma^2_{\rm res}(x_i)$ is only half $\sigma^2_{\rm diff}(x_i)$.

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

$$\sigma^2(l) = \sigma^2(x_2) - 2{\rm cov}(x_2, x_1) + \sigma^2(x_1). \tag{16}$$

For the diffraction data alone, this gives

$$\sigma^2_{\rm diff}(l) = 1/a + 0 + 1/a = 2/a = 2\sigma^2_{\rm diff}(x_i), \tag{17}$$

as expected. For the diffraction data plus restraint,

$$\begin{aligned} \sigma^2_{\rm res}(l) &= [1/a(a+2b)][(a+b) - 2b + (a+b)] \\ &= 1/(a/2+b) \\ &< \sigma^2_{\rm diff}(l). \end{aligned} \tag{18}$$

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), can be re-expressed as

$$1/\sigma^2_{\rm res}(l) = 1/\sigma^2_{\rm diff}(l) + 1/\sigma^2_{\rm geom}(l). \tag{19}$$

This form proves very useful in the real examples considered later.

For the two-atom protein it can be proved directly, as one would expect from (19), *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 examples of the s.u.s in restrained refinement, suppose the stereochemical length restraint has $\sigma_{\rm geom}(l)$ = 0.02 Å. Equation (18) gives the length s.u. $\sigma_{\rm res}(l)$ in restrained refinement. If the diffraction-only $\sigma_{\rm diff}(x_i)$ = 0.01, 0.02 or 0.05 Å, the restrained $\sigma_{\rm res}(l)$ is 0.012, 0.016 or 0.019 Å, respectively. However large $\sigma_{\rm diff}(x_i)$, $\sigma_{\rm res}(l)$ never exceeds 0.02 Å.

Equation (15) gives the position s.u. $\sigma_{\rm res}(x_i)$ in restrained refinement. If the diffraction-only $\sigma_{\rm diff}(x_i)$ = 0.01, 0.02 or 0.05 Å, the restrained $\sigma_{\rm res}(x_i)$ is 0.009, 0.016 or 0.037 Å, respectively. 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.

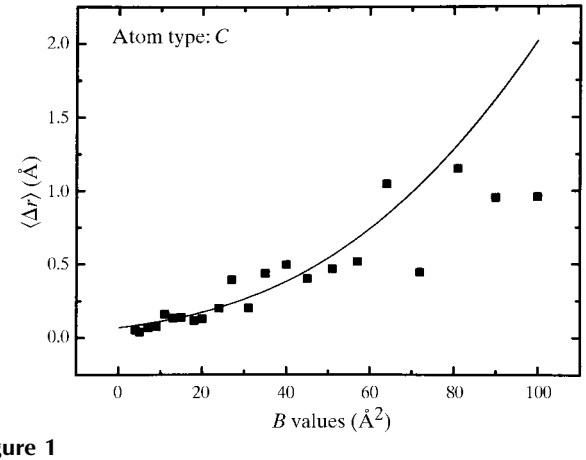

Figure 1
Comparison of the r.m.s. position differences ⟨Δ*r*⟩ between the C atoms in the structures of 1TGI and 1TGF and the theoretical distribution curve derived from the error formula (1). Reproduced from Daopin *et al.* (1994).

4. Examples of full-matrix inversion

4.1. Unrestrained and restrained inversions for concanavalin A

G. M. Sheldrick has kindly 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 with data to 0.94 Å refined anisotropically at 110 K. 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 × 3) are plotted in Figs. 2, 3 and 4. Fig. 2 shows $\sigma(r)$ *versus* B_{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

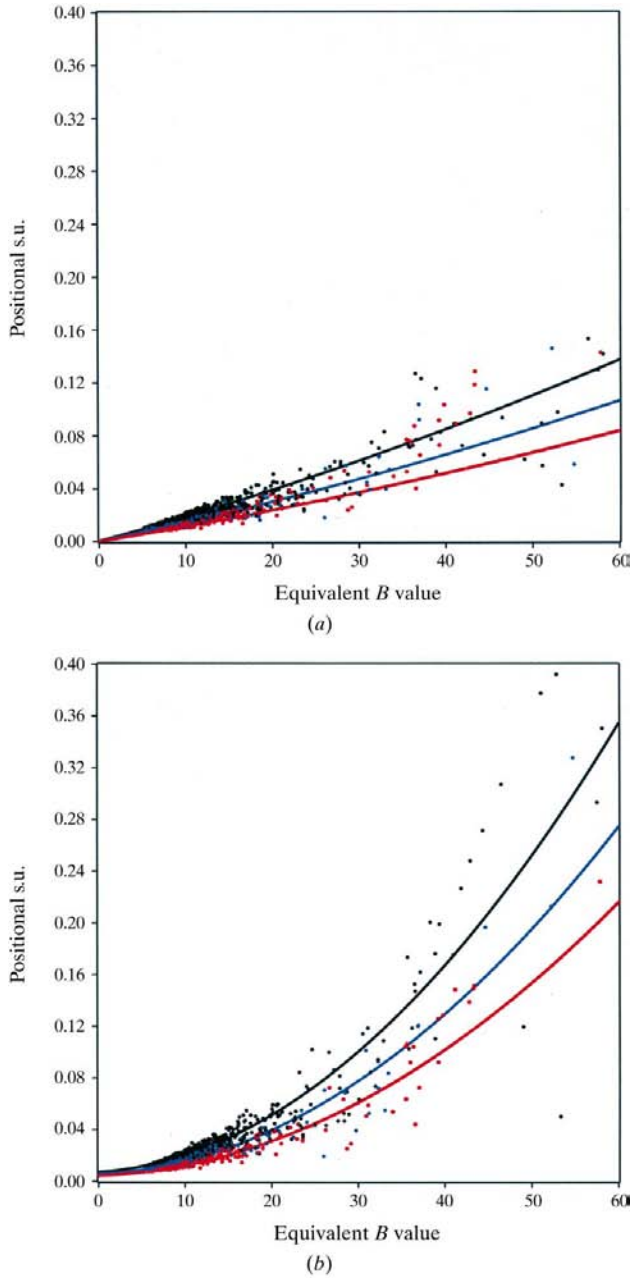

Figure 2
Plots of $\sigma(r)$ *versus* B_{eq} for concanavalin A with 0.94 Å data, (*a*) restrained full-matrix $\sigma_{res}(r)$, (*b*) unrestrained full-matrix $\sigma_{diff}(r)$. Carbon black, nitrogen blue, oxygen red.

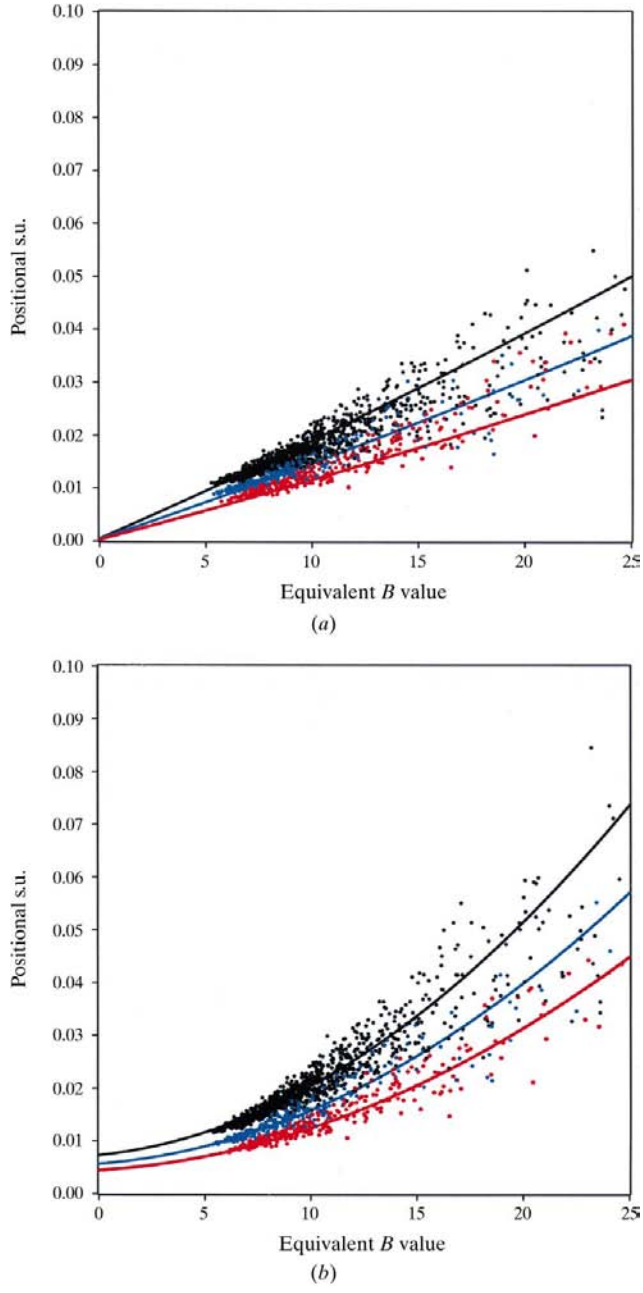

Figure 3
Plots for low B of $\sigma(r)$ *versus* B_{eq} for concanavalin A with 0.94 Å data, (*a*) restrained full-matrix $\sigma_{res}(r)$, (*b*) unrestrained full-matrix $\sigma_{diff}(r)$. Carbon black, nitrogen blue, oxygen red.

for nitrogen and red for oxygen. Fig. 2(*a*) shows the restrained results and Fig. 2(*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 $\sigma(r)$ fall below 0.02 Å and both are around 0.01 Å at $B = 6$ Å^2.

The behaviour of $\sigma(r)$ at low B is shown in more detail in Fig. 3. For $B < 10$ Å^2, the better precision of oxygen compared with nitrogen and of nitrogen compared with carbon can be clearly seen. At the lowest B, the unrestrained $\sigma_{\rm diff}(r)$ in Fig. 3(*b*) are almost as small as the restrained $\sigma_{\rm res}(r)$ in Fig. 3(*a*). [The quadratic fits of the restrained results in Figs. 2(*a*) and 3(*a*) are evidently slightly imperfect in making $\sigma_{\rm res}(r)$ tend almost to 0 as B tends to 0.]

Fig. 4 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. 4(*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. 4(*b*) goes off the scale of the diagram. However, 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 §3.2. For that model, the equation

$$1/\sigma^2_{\rm res}(l) = 1/\sigma^2_{\rm diff}(l) + 1/\sigma^2_{\rm geom}(l) \qquad (19)$$

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 (19) with the unrestrained $\sigma_{\rm diff}(l)$ of any bond, the predicted restrained $\sigma_{\rm res}(l)$ is close to that found from the restrained full matrix (Fig. 5).

It can be seen from Fig. 4(*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 $\sigma_{\rm res}(l)$ 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 $\sigma_{\rm diff}(l)$ (*i.e.* high B), equation (19) predicts that $\sigma_{\rm res}(l) = \sigma_{\rm geom}(l) = 0.02$ Å, as is found in Fig. 4(*a*).

In an isotropic approximation $\sigma(r) = 3^{1/2}\sigma(x)$. In the two-atom model, (15) can be rearranged in the form

$$\sigma^2_{\rm res}(x) = \sigma^2_{\rm diff}(x)\{[\sigma^2_{\rm diff}(x) + 0.02^2]/[2\sigma^2_{\rm diff}(x) + 0.02^2]\}. \qquad (20)$$

To derive a predicted value of $\sigma_{\rm res}(r)$ from a value of $\sigma_{\rm diff}(r)$, one must divide $\sigma_{\rm diff}(r)$ by $3^{1/2}$ to obtain $\sigma_{\rm diff}(x)$, use (20) to obtain $\sigma_{\rm res}(x)$ and then multiply by $3^{1/2}$ to obtain $\sigma_{\rm res}(r)$.

For low B, say $B \leq 15$ Å^2 in concanavalin, (20) gives quite good predictions of $\sigma_{\rm res}(r)$ from $\sigma_{\rm diff}(r)$. For instance, for a C atom with $B = 15$ Å^2, the quadratic curve for carbon in Fig. 3(*b*) shows $\sigma_{\rm diff}(r) = 0.034$ Å and Fig. 3(*a*) shows $\sigma_{\rm res}(r) = 0.029$ Å. If $\sigma_{\rm diff}(r) = 0.034$ Å $= 3^{1/2}\sigma_{\rm diff}(x)$ is used with (20), the resulting prediction for $\sigma_{\rm res}(r)$ is 0.028 Å.

However, for high B, say $B = 50$ Å^2, the quadratic curve for carbon in Fig. 2(*b*) shows $\sigma_{\rm diff}(r) = 0.25$ Å and Fig. 2(*a*) shows $\sigma_{\rm res}(r) = 0.11$ Å, whereas (20) leads to the poor estimate $\sigma_{\rm res}(r) = 0.18$ Å.

Thus, at high B, equation (20) from the two-atom model does not give a good description of the relation 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 O atom linked to its C atom by a

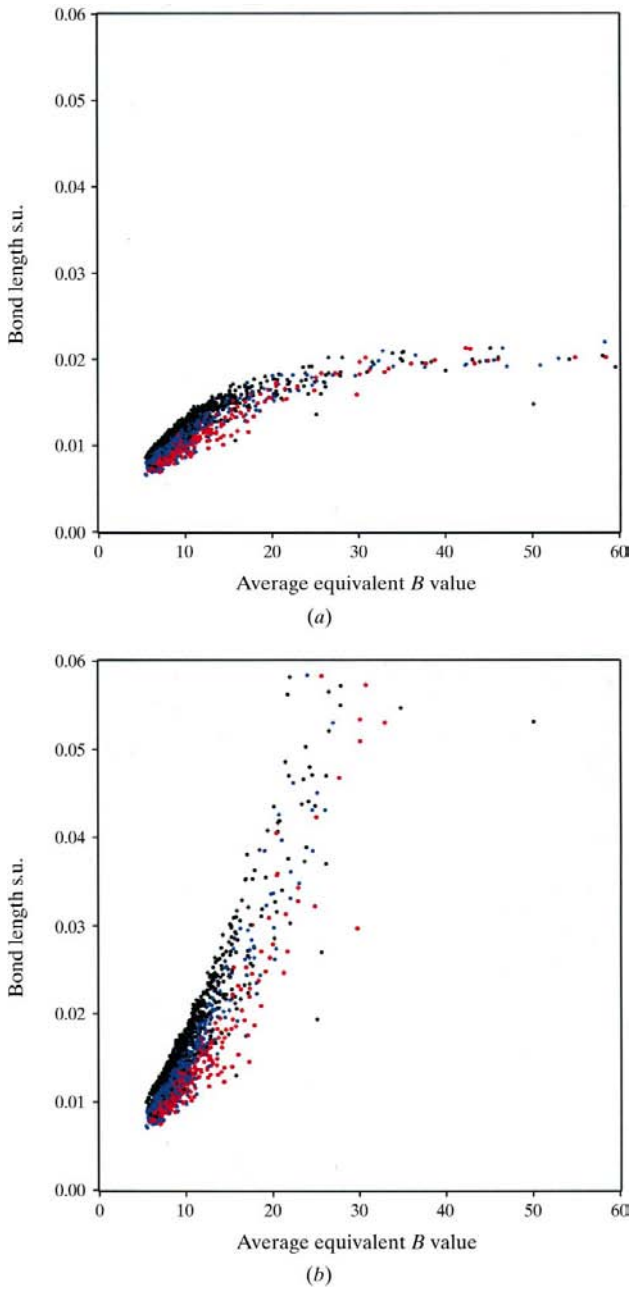

Figure 4
Plots of $\sigma(l)$ *versus* average $B_{\rm eq}$ for concanavalin A with 0.94 Å data, (*a*) restrained full-matrix $\sigma_{\rm res}(l)$, (*b*) unrestrained full-matrix $\sigma_{\rm diff}(l)$. C—C black, C—N blue, C—O red.

0.02 Å restraint is also subject to 0.04 Å 1–3 restraints to chain Cα 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 $\sigma_{res}(r)$ is lower, compared with the diffraction-data-only $\sigma_{diff}(r)$, by a greater amount than would be expected from the two-atom model.

4.2. Unrestrained inversions for cytochrome *c*6 and immunoglobulin

Sheldrick has also kindly provided the results of unrestrained inversions for two proteins: (i) a cytochrome *c*6 with 89 residues, with data to 1.10 Å refined anisotropically (Frazão *et al.*, 1995); (ii) a single-chain immunoglobulin mutant (t39k) with 218 amino-acid residues, with data to 1.70 Å refined isotropically (Usón *et al.*, 1999). Figs. 6(*a*) and 6(*b*) show $\sigma_{diff}(r)$ *versus* B_{eq} for the fully occupied protein atoms in the cytochrome *c*6 and in the immunoglobulin. As expected, the cytochrome results are more precise. Superposed on the data points are least-squares quadratic fits. In the very rough approximation for $\sigma_{diff}(x_i)$ suggested later by equation (24), the dependence on atom type was controlled by $N_i^{1/2} = (\sum Z_j^2/Z_i^2)^{1/2}$. Sheldrick found that a $Z_i^{\#}$ dependence produced too little difference between C, N and O. The quadratics for $\sigma(r)$ in the figures are based on Z_i, the scattering factors at $\sin\theta/\lambda = 0.3$ Å^{-1}. For C, N and O these are 2.494, 3.219 and 4.089, respectively. For potential use in §6 or in subsequent work, the least-squares fits to the $Z_i^{\#}$, $\sigma(r_i)$ in Å are recorded here as

$$0.01512 + 0.001778B + 0.0001452B^2, \quad (21a)$$

$$0.11892 + 0.008910B + 0.0001462B^2, \quad (21b)$$

$$0.01826 + 0.001043B + 0.0002230B^2, \quad (21c)$$

$$0.00115 + 0.004414B + 0.0000214B^2, \quad (21d)$$

for cytochrome *c*6 (unrestrained), immunoglobulin (unrestrained), concanavalin A (unrestrained) and concanavalin A (restrained), respectively.

Figs. 7(*a*) and 7(*b*) show $\sigma_{diff}(l)$ *versus* B_{eq} for the cytochrome and immunoglobulin. Note that the lowest immunoglobulin unrestrained $\sigma_{diff}(l)$ is about 0.06 Å, which is three times the 0.02 Å $\sigma_{geom}(l)$ bond restraint. For cytochrome *c*6 a

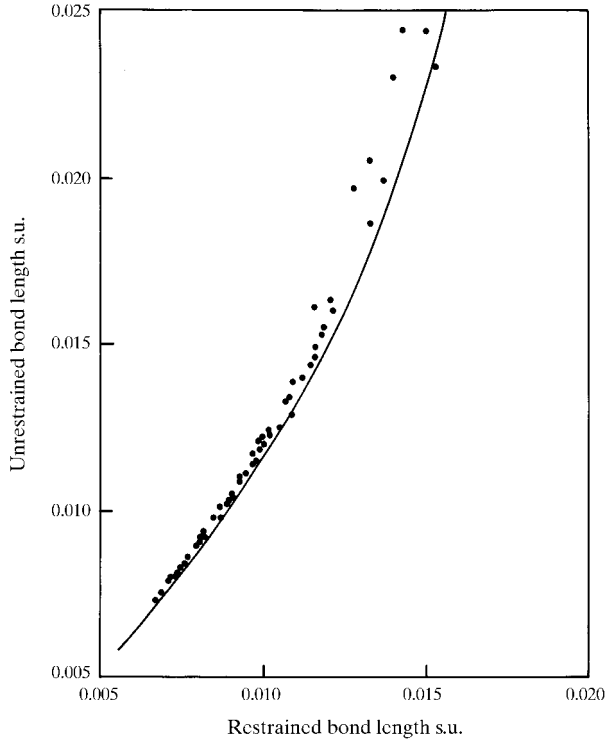

Figure 5
Unrestrained $\sigma_{diff}(l)$ *versus* restrained $\sigma_{res}(l)$ for some bonds in concanavalin A. The continuous-line curve is from the two-atom protein model equation (19) with length restraint $\sigma_{geom}(l) = 0.02$ Å.

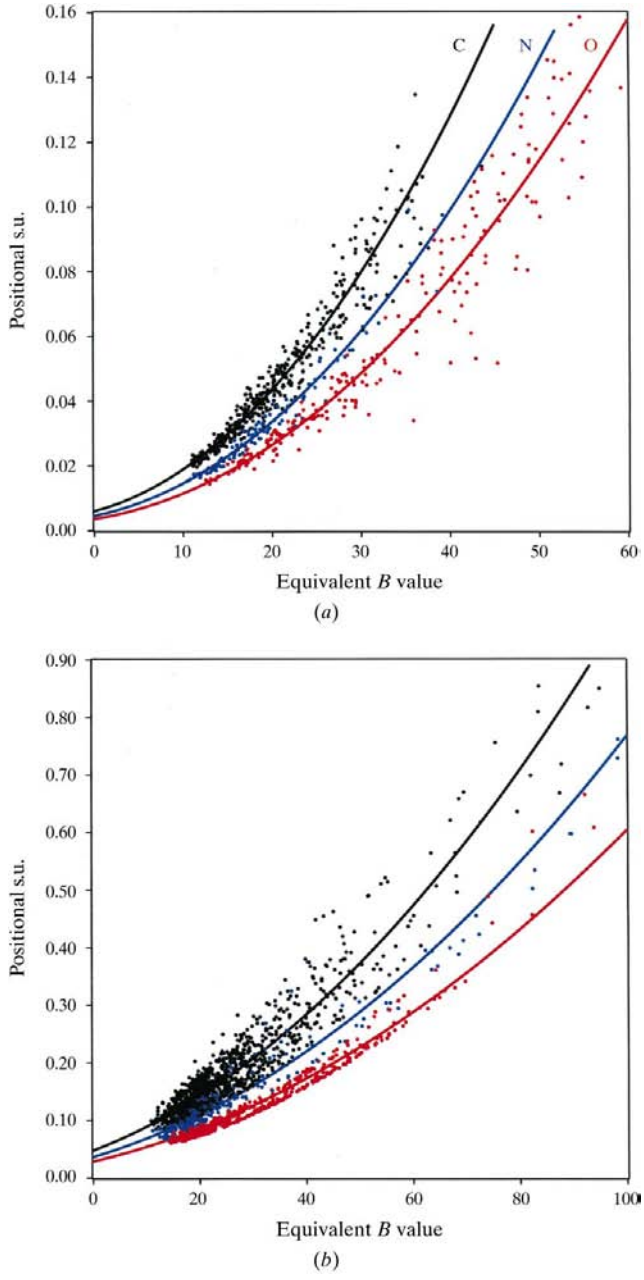

Figure 6
Plots of $\sigma_{diff}(r)$ *versus* B_{eq} from unrestrained full matrix, (*a*) cytochrome *c*6 with 1.10 Å data, (*b*) immunoglobulin mutant (t39k) with 1.70 Å data. Carbon black, nitrogen blue, oxygen red.

few $\sigma_{diff}(l)$ are below the 0.02 Å level, whereas for unrestrained concanavalin many $\sigma_{diff}(l)$ are below that level.

5. Approximate methods

The full-matrix inversions described in the last section required massive calculations. The length of the calculations is more a matter of the order of the matrix, *i.e.* number of parameters, than of the number of observations. When restraints have been applied, it is the diffraction-cum-restraints full matrix which should be inverted. If the full calculation is deemed uneconomic or impracticable, there are a variety of possible simplifications of increasing simplicity and roughness.

As long ago 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 blocks.

An extreme example of an apparently successful gross approximation is represented by the treatment of TGF-β2 by Daopin *et al.* (1994) discussed above in §2 and illustrated in Fig. 1. The Fourier map formula (Cruickshank, 1949*a*, 1952, 1959),

$$\sigma(x) = \sigma(\text{slope})/(\text{atomic peak 'curvature'}), \qquad (1)$$

yielded a quite good description of the B dependence of the differences between two independent determinations of the same protein. However, there is a formal difficulty about this application. Equation (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, corresponding to the b matrix (9) of the two-atom model, it is somewhat surprising that (1) should have worked well.

It is less surprising that Chambers & Stroud considered that (1) underestimated the errors. The two structures (Chambers & Stroud, 1979; Bode & Schwager, 1975) of bovine trypsin which they compared were refined by different methods and they seem to indicate that the refinements had not reached convergence.

Equation (1) requires the summation of various series over all (hkl) observations; such calculations are not customarily provided in protein programs. However, owing to the fundamental similarities between Fourier and least-squares methods demonstrated by Cochran (1948), Cruickshank (1949*b*) and Cruickshank & Robertson (1953) and outlined in *Appendix* §A.2, 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 already have been calculated 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 off-diagonal terms could be suggested. Computational trials would be needed to explore whether satisfactory compromises can be found which avoid the massive calculations of full matrices.

An especially simple and gross approximation for diffraction-only data will now be considered.

6. The diffraction-component precision index (DPI)

6.1. Statistical expectation of error dependence

From general statistical theory, one would expect the s.u. of an atomic coordinate determined from the diffraction data alone to show dependence on four factors:

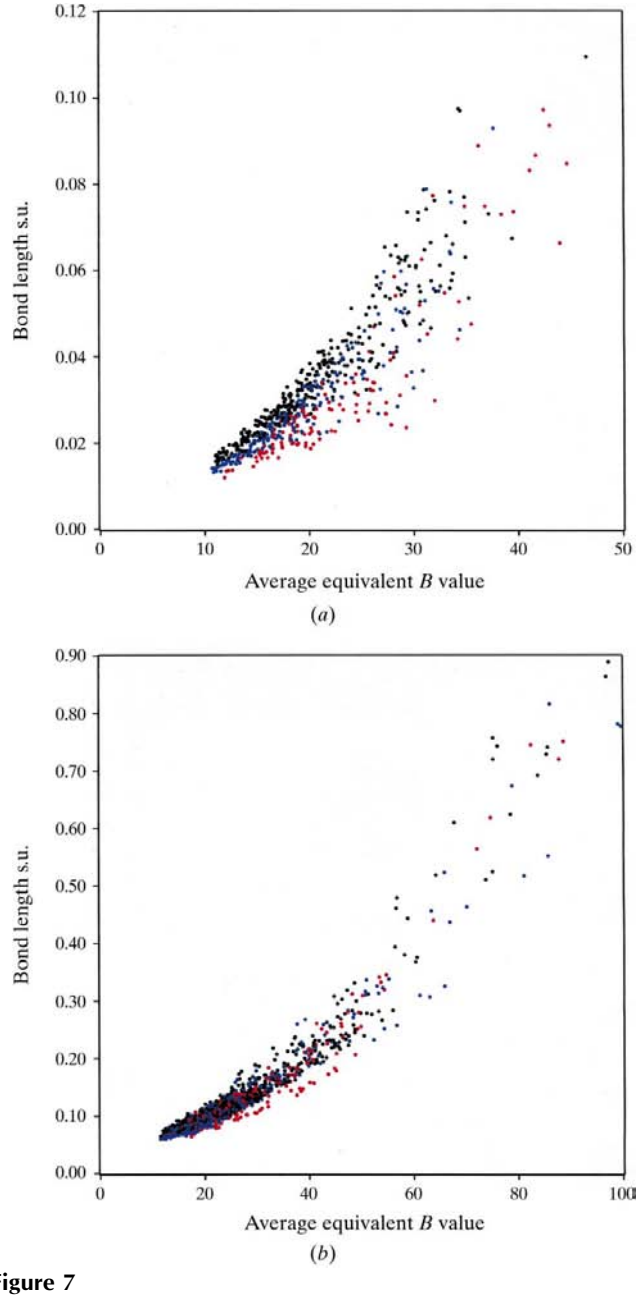

Figure 7
Plots of $\sigma_{diff}(2)$ *versus* B_{eq} from unrestrained full matrix (*a*) cytochrome *c*6 with 1.10 Å data, (*b*) immunoglobulin mutant (t39k) with 1.70 Å data. Carbon black, nitrogen blue, oxygen red.

$$\sigma(x) \propto (\mathcal{R})[(n_{\rm atoms})/(n_{\rm obs} - n_{\rm params})]^{1/2}(1/s_{\rm rms}). \quad (22)$$

Here, $\mathcal{R}$ is some measure of the precision of the data, $n_{\rm atoms}$ is the recognition that the information content of the data has to be shared out, $n_{\rm obs}$ is the number of independent data (but to achieve the correct number of degrees of freedom this must be reduced by $n_{\rm params}$, the number of parameters determined) and $1/s_{\rm rms}$ is a more specialized factor arising from the sensitivity $\partial|F|/\partial x$ of the data to the parameter x. Here, $s_{\rm rms}$ is the r.m.s. reciprocal radius of the data. Any statistical error estimate must show some correspondence to these four factors.

6.2. A simple error formula

Cruickshank (1960), based on a least-squares approach (see *Appendix* §A.1), offered a simple order-of-magnitude formula for $\sigma(x)$ in small molecules. It was intended for use in experimental design: how many data of what precision are needed to achieve a given precision in the results? The formula, derived from a very rough estimate of a least-squares diagonal element in non-centrosymmetric space groups, was

$$\sigma(x_i) = (1/2)(N_i/p)^{1/2}(R/s_{\rm rms}). \quad (23)$$

Here, $p = n_{\rm obs} - n_{\rm params}$, R is the usual residual $\sum|\Delta F|/\sum|F|$ and N_i is the number of atoms of type i needed to give scattering power at $s_{\rm rms}$ equal to that of the asymmetric unit of the structure, *i.e.* $\sum_j f_j^2 \equiv N_i f_i^2$. [The formula has also proved very useful in a systematic study of coordinate precision in the many thousands of small-molecule structure analyses recorded in the Cambridge Structural Database (Allen *et al.*, 1995*a*,*b*).]

For small molecules, the above definition of N_i allowed the treatment of different types of atom with not dissimilar B values. However, it is not suitable for individual atoms in proteins where there is a very large range of B values and some atoms have B values so large as to possess negligible scattering power at $s_{\rm rms}$.

Often, as in isotropic refinement, $n_{\rm params} \simeq 4n_{\rm atoms}$, where $n_{\rm atoms}$ is the total number of atoms in the asymmetric unit. For fully anisotropic refinement $n_{\rm params} \simeq 9n_{\rm atoms}$.

A first very rough extension of (23) for application in proteins to an atom with $B = B_i$ is

$$\sigma(x_i) = k(N_i/p)^{1/2}[g(B_i)/g(B_{\rm avg})]C^{-1/3}Rd_{\rm min}, \quad (24)$$

where k is about 1.0, $N_i = \sum Z_j^2/Z_i^2$, $B_{\rm avg}$ is the average B for fully occupied sites and C is the fractional completeness of the data to $d_{\rm min}$. In deriving (24) from (23), $1/s_{\rm rms}$ has been replaced by $1.3d_{\rm min}$ and the factor $(1/2)(1.3) = 0.65$ has been increased to 1.0 as a measure of caution in the replacement of a full matrix by a diagonal approximation. $g(B) = 1 + a_1B + a_2B^2$ is an empirical function to allow for the dependence of $\sigma(x)$ on B. However, the results in (21) show that the parameters a_1 and a_2 depend upon the structure. (See Stroud & Fauman, 1995 for the form of possible three-parameter exponential functions.)

As already mentioned, Sheldrick has found that the Z_i in N_i is better replaced by $Z_i^{\#}$, the scattering factor at $\sin\theta/\lambda = 0.3$ Å^{-1}. Hence, N_i may be taken as

$$N_i = \left(\sum Z_j^{\#2}/Z_i^{\#2}\right). \quad (25)$$

A useful comparison of the relative precision of different structures may be obtained by comparing atoms with the respective $B = B_{\rm avg}$ in the different structures. (24) then reduces to

$$\sigma(x, B_{\rm avg}) = 1.0(N_i/p)^{1/2}C^{-1/3}Rd_{\rm min}. \quad (26)$$

The smaller $d_{\rm min}$ and R, the better the precision of the structure. If the difference between O, N and C atoms is ignored, N_i may be taken simply as the number of fully occupied sites. For heavy atoms, (25) must be used for N_i.

Equation (26) is not to be regarded as having absolute validity. It is a quick and rough guide for the diffraction-data-only error component for an atom with Debye B equal to the $B_{\rm avg}$ for the structure. We shall call it the **diffraction-component precision index** (DPI). It contains none of the restraint data.

6.3. Extension for low-resolution structures and use of $R_{\rm free}$

For low-resolution structures, the number of parameters may exceed the number of diffraction data. In (24) and (26) $p = n_{\rm obs} - n_{\rm params}$ is then negative, so that $\sigma(x)$ is imaginary. This difficulty can be circumvented empirically by replacing p with $n_{\rm obs}$ and R with $R_{\rm free}$ (Brünger, 1992). The counterpart of the DPI (26) is then

$$\sigma(x, B_{\rm avg}) = 1.0(N_i/n_{\rm obs})^{1/2}C^{-1/3}R_{\rm free}d_{\rm min}. \quad (27)$$

Here, $n_{\rm obs}$ is the number of reflections included in the refinement, not the number in the $R_{\rm free}$ set.

It may be asked: how can there be any estimate for the precision of a coordinate from the diffraction data only when there is insufficient diffraction data to determine the structure? By following the line of argument of Cruickshank's (1960) analysis, (27) is a rough estimate of the square root of the reciprocal of one diagonal element of the diffraction-only least-squares matrix. All the other parameters can be regarded as having been determined from a diffraction-plus-restraints matrix (see also the discussion in §9).

Clearly, (27) can also be used as a general alternative to (26) as a diffraction-component precision index (DPI), irrespective of whether the number of degrees of freedom $p = n_{\rm obs} - n_{\rm params}$ is positive or negative.

6.3.1. Comment. When p is positive, (27) would be exactly equivalent to (26) only if $R_{\rm free} = R[n_{\rm obs}/(n_{\rm obs} - n_{\rm params})]^{1/2}$. Tickle *et al.* (1998*b*) have shown that the expected relation in a restrained refinement is actually

$$R_{\rm free} = R\{[n_{\rm obs} + (n_{\rm params} - h)]/[n_{\rm obs} - (n_{\rm params} - h)]\}^{1/2}, \quad (28)$$

where $h = n_{\rm restraints} - \sum w_{\rm geom}(\Delta Q)^2$, the latter term, as in (4), being the weighted sum of the squares of the restraint residuals.

6.4. Position error

Often an estimate of a position error $|\Delta\mathbf{r}|$ is required rather than that of a coordinate error $|\Delta x|$. In the isotropic approximation

$$\sigma(r, B_{\rm avg}) = 3^{1/2}\sigma(x, B_{\rm avg}). \quad (29)$$

Consequently the DPI formulae for the position errors are

$$\sigma(r, B_{\rm avg}) = 3^{1/2}(N_i/p)^{1/3}C^{-1/2}Rd_{\rm min} \quad (30)$$

with R and

$$\sigma(r, B_{\rm avg}) = 3^{1/2}(N_i/n_{\rm obs})^{1/2}C^{-1/3}R_{\rm free}d_{\rm min} \quad (31)$$

with $R_{\rm free}$.

7. Examples of the diffraction-component precision index

7.1. Full-matrix comparison with the diffraction-component precision index

The DPI (30) with R was offered as a quick and rough guide for the diffraction-data-only error for an atom with $B = B_{\rm avg}$. The necessary data for the comparisons with the three unrestrained full-matrix inversions of §4 are given in Table 1. For concanavalin with $B_{\rm avg} = 14.8$ Å^2, the full-matrix quadratic (21*c*) gives 0.033 Å for a C atom and the DPI gives 0.034 Å for an unspecified atom. For cytochrome with $B_{\rm avg} = 24.0$ Å^2, the full-matrix quadratic (21*a*) gives 0.057 Å for a C atom and the DPI gives 0.066 Å for $\sigma(r,B_{\rm avg})$. For immunoglobulin with $B_{\rm avg} = 26.8$ Å^2, the full-matrix quadratic (21*b*) gives $\sigma_{\rm diff}(r) = 0.19$ Å for a C atom, while the DPI gives 0.22 Å.

For these three structures the 'back-of-an-envelope' DPI formula (30) compares remarkably well with the full-matrix calculations at $B_{\rm avg}$.

For the restrained full-matrix calculations on concanavalin A, the quadratic (21*d*) at $B_{\rm avg}$ gives for a C atom $\sigma_{\rm res}(r) = 0.028$ Å, which is only 15% smaller than the unrestrained 0.033 Å. This small decrease matches the discussion of $\sigma_{\rm res}(r)$ and $\sigma_{\rm diff}(r)$ in §4.1. But that discussion also indicates that for the immunoglobulin the restrained $\sigma_{\rm res}(r,B_{\rm avg})$, which was not computed, will be proportionately much lower than the unrestrained value of $\sigma_{\rm diff}(r,B_{\rm avg}) = 0.19$ Å since the restraints are relatively more important in the immunoglobulin.

7.2. Further examples of the DPI using *R*

Table 2 shows some examples of the application of the diffraction-component precision index (30) with R to proteins of differing precision, starting with the smallest $d_{\rm min}$. In all the examples N_i has been set equal to $n_{\rm atoms}$, the total number of atoms. The right-hand columns show $\langle\Delta r\rangle$ values derived from Luzzati (1952) and Read (1986) plots described later in §8.

The first entry is for crambin at 0.83 Å resolution and 130 K (Stec, Zhou *et al.*, 1995). Their results were obtained from an unrestrained full-matrix anisotropic refinement. Inversion of the full matrix gave $\sigma_{\rm diff}(x)$ s.u.s of 0.0096 Å for backbone atoms, 0.0168 Å for side-chain atoms and 0.0409 Å for solvent atoms, with an average for all atoms of 0.022 Å. The DPI $\sigma(r,B_{\rm avg}) = 0.021$ Å corresponds to $\sigma(x) = 0.012$ Å, which is satisfactorily intermediate between the full-matrix values for the backbone and side-chain atoms.

Table 1
Comparison of full-matrix $\sigma(r,B_{\rm avg})$ with diffraction-data precision indicator DPI.

Protein	$(N_i/p)^{1/2}$	R	$d_{\rm min}$ (Å)	DPI $\sigma(r,B_{\rm avg})$ (Å)	Full matrix $\sigma_{\rm diff}(r,B_{\rm avg})$ (Å)	References
Concanavalin A	0.148	0.128	0.94	0.034	0.033	*a*
Cytochrome *c*6	0.244	0.140	1.10	0.066	0.057	*b*
Immunoglobulin	0.476	0.156	1.70	0.221	0.186	*c*

References: (*a*) Deacon *et al.* (1997); (*b*) Frazão *et al.* (1995); (*c*) Usón *et al.* (1999).

The next entry is for rubredoxin at 1.0 Å (Dauter *et al.*, 1992). They carried out both unrestrained and restrained isotropic full-matrix refinements. Details are given in Table 2 for the unrestrained refinement. They did not make formal calculations of s.u.s, but from the deviations of the bond lengths from the dictionary values, they suggested the r.m.s errors in the coordinates of the well ordered atoms were about 0.04 Å. The DPI corresponds to $\sigma(x,B_{\rm avg}) = 0.028$ Å.

Sevcik *et al.* (1996) carried restrained anisotropic full-matrix refinements on data from two slightly different crystals of ribonuclease Sa with $d_{\rm min}$ of 1.15 and 1.20 Å. They inverted full-matrix blocks containing parameters of 20 residues to estimate coordinate errors. The overall r.m.s. coordinate error for protein atoms is given as 0.03 Å and for all atoms (including waters and ligands) as 0.07 Å for MGMP and 0.05 Å for MSA. The DPI gives $\sigma(r,B_{\rm avg}) = 0.05$ Å for both structures.

The next entries concern poplar plastocyanin at 295 K (Guss *et al.*, 1992) and at 173 K (Fields *et al.*, 1994). For the 173 K study, a single set of Hamburg synchrotron data was refined quite independently with different programs in Sydney and Hamburg. The r.m.s. difference in position of the protein atoms between the two models from the same data was 0.12 Å (excluding six outliers). As would be hoped, this is less than the 0.21 and 0.24 Å DPI of the individual refinements. The higher resolution 295 K study has a smaller DPI of 0.11 Å because it has twice as many data.

The next entries concern the two lower resolution studies of TGF-β2 (Daopin *et al.*, 1994). The DPI gives $\sigma(r) = 0.16$ Å for 1TGI and 0.24 Å for 1TGF. This indicates an r.m.s. position difference between the structures for atoms with $B_i = B_{\rm avg}$ of $(0.16^2 + 0.24^2)^{1/2} = 0.29$ Å. Daopin *et al.* (1994) reported the differences between the two determinations, omitting poor parts, as $\langle\Delta r\rangle_{\rm r.m.s.} = 0.15$ Å (main chain) and 0.29 Å (all atoms).

Cadmium-azurin (Blackwell *et al.*, 1994) with $d_{\rm min} = 1.8$ Å gives a DPI $\sigma(r,B_{\rm avg}) = 0.21$ Å. Human differic lactoferrin (Haridas *et al.*, 1995) is an example of a large protein at the lower resolution of 2.2 Å with a high value of $(N_i/p)^{1/2}$ leading to $\sigma(r,B_{\rm avg}) = 0.43$ Å.

Three crystal forms of thaumatin were studied by Ko *et al.* (1994). The orthorhombic and tetragonal forms diffracted to

Table 2
Examples of diffraction-component precision index DPI.

Protein	N_i	n_{obs}	$(N_i/p)^{1/2}$	$C^{-1/3}$	R	d_{min} (Å)	DPI $\sigma(r,B_{avg})$ (Å)	Luzzati $\langle\Delta r\rangle$ (Å)	Read $\langle\Delta r\rangle$ (Å)	Reference
Crambin	447	23759	0.150	1.074	0.090	0.83	0.021	0.055		Stec, Zhou *et al.* (1995)
Rubredoxin	479	18532	0.170	1.034	0.160	1.00	0.049	∼0.13	0.13	Dauter *et al.* (1992)
Ribonuclease MGMP	1958	62845	0.208	1.046	0.109	1.15	0.047		0.08	Sevcik *et al.* (1996)
Ribonuclease MSA	1832	60670	0.204	1.016	0.106	1.20	0.045		0.05	Sevcik *et al.* (1996)
Plastocyanin, 295 K	849	14303	0.279	1.096	0.149	1.33	0.11	0.15		Guss *et al.* (1992)
Plastocyanin, 173 K, Sydney	928	7393	0.502	1.13	0.132	1.60	0.21	0.13		Fields *et al.* (1994)
Plastocyanin, 173 K, Hamburg	911	7393	0.493	1.13	0.153	1.60	0.24			Fields *et al.* (1994)
TGF-β2, 1TGI	948	∼14000	0.305	∼1.0	0.173	1.80	0.16	∼0.20	0.18	Daopin *et al.* (1994)
TGF-β2, 1TFG	974	∼11000	0.370	∼1.0	0.188	1.95	0.24	∼0.24		Daopin *et al.* (1994)
Cd-azurin	2215	23449	0.391	1.02	0.168	1.80	0.21	0.15	0.24	Blackwell *et al.* (1994)
Lactoferrin	5907	39113	0.618	1.036	0.179	2.20	0.43	0.25–0.30	0.35	Haridas *et al.* (1995)
Thaumatin C2	1552	4622	†	1.10	0.184	2.60	—	0.25		Ko *et al.* (1994)

† (N_i/p) negative.

1.75 Å, but the monoclinic *C*2 form diffracted only to 2.6 Å. The structures with 1552 protein atoms were successfully refined with restraints by *X-PLOR* and *TNT*. For the monoclinic form, the number of parameters exceeds the number of diffraction observations, so (N_i/p) is negative and no estimate by (30) of the diffraction-data-only error is possible. The DPI (30) gives 0.17 and 0.16 Å for the orthorhombic and tetragonal forms.

Peters-Libeu & Adman (1997) recently studied the structural differences between oxidation states of several pseudoazurins and also between several plastocyanins. The main method they devised was named a displacement-parameter weighted coordinate comparison. A preliminary version of the DPI (Dodson *et al.*, 1996), which they also used as an indicator, contained a factor 0.7 rather than the 1.0 now shown in (26).

7.3. Examples of the DPI using R_{free}

As in the case of monoclinic thaumatin, for low-resolution structures the number of parameters may exceed the number of diffraction data. To circumvent this difficulty, it was proposed in §6.3 to replace $p = n_{obs} - n_{params}$ by n_{obs} and R by R_{free} in a revised formula (31) for the DPI. Table 3 shows examples for some structures for which both R and R_{free} were available. The second line of each structure shows the alternative values for $(N_i/p)^{1/2}$, R_{free} and the DPI $\sigma(r,B_{avg})$ from (31).

It will be seen that for the structures with $d_{min} \leq 2.0$ Å, the DPI is much the same whether it is based on R or R_{free}.

Tickle *et al.* (1998*a*) have made full-matrix error estimates for isotropic restrained refinements of γB-crystallin with d_{min} = 1.49 Å and βB2-crystallin with d_{min} = 2.10 Å. The DPI $\sigma(r,B_{avg})$ are calculated for the two structures as 0.14 and 0.25 Å, respectively, with R in (30), and as 0.14 and 0.22 Å, respectively, with R_{free} in (31). The full-matrix weighted averages of $\sigma_{res}(r)$ for all protein atoms were 0.10 and 0.15 Å, respectively; for only main-chain atoms 0.05 and 0.08 Å, for side-chain atoms 0.14 and 0.20 Å and for water O atoms 0.27 and 0.35 Å, respectively. Again, the DPI gives reasonable overall indices of the quality of the structures.

For the complex of bovine ribonuclease A and porcine ribonuclease inhibitor (Kobe & Deisenhofer, 1995) with d_{min} = 2.50, the number of reflections is only just larger than the number of parameters, so that $(N_i/p)^{1/2}$ = 1.922 is very large and the DPI with R gives an unrealistic 1.85 Å. With R_{free}, $\sigma(r,B_{avg})$ = 0.69 Å.

The HyHEL-5–lysozyme complex (Cohen *et al.*, 1996) had d_{min} = 2.65 Å. Here, the number of reflections is fewer than the number of parameters, but the R_{free} formula gives $\sigma(r,B_{avg})$ = 0.69 Å.

Table 3 will be considered further in §8 as part of the discussion of the Luzzati and Read methods for $\langle\Delta r\rangle$.

8. Critique of Luzzati plots

8.1. Luzzati's theory

Luzzati (1952) did not provide a theory for estimating positional errors at the end of a normal refinement. He provided a theory for estimating the positional changes needed in a further idealized refinement to reach R = 0.

(i) His theory assumed that the F_{obs} had no errors, and that the F_{calc} model (scattering factors, thermal parameters *etc.*) was perfect apart from coordinate errors.

(ii) The Gaussian probability distribution for these coordinate errors was assumed to be the *same for all atoms*, independent of Z or B.

(iii) The atoms were not required to be identical, and the position errors were not required to be small.

Luzzati gave families of curves for R *versus* $\sin\theta/\lambda$ for varying average positional errors $\langle\Delta r\rangle$ for both centrosymmetric and non-centrosymmetric structures. The curves do not depend on the number N of atoms in the cell. They all rise from R = 0 at $\sin\theta/\lambda$ = 0 to the Wilson (1950) values 0.828 and 0.586 for random structures at high $\sin\theta/\lambda$. In a footnote (p. 807) Luzzati suggested that at the end of a normal refinement (with R non-zero owing to experimental and model errors, *etc.*) the curves would indicate an upper limit for $\langle\Delta r\rangle$. He noted that typical small-molecule $\sigma(r)$ of 0.01–0.02 Å, if used as $\langle\Delta r\rangle$ in the plots, would give much smaller R than are found at the end of a refinement.

Table 3
Comparison of DPI using R and R_{free}.

The second row for each protein contains values appropriate to the DPI equation (31) using R_{free}.

Protein	N_i	n_{obs}	$(N_i/p)^{1/2}$ $(N_i/n_{obs})^{1/2}$	$C^{-1/3}$	R R_{free}	d_{min} (Å)	DPI $\sigma(r,B_{avg})$ (Å)	Luzzati $\langle\Delta r\rangle$ (Å)	Read $\langle\Delta r\rangle$ (Å)	Reference
Concanavalin A	2130	116712	0.148	1.099	0.128	0.94	0.034	0.06		Deacon *et al.* (1997)
			0.135		0.148		0.036			
HEW lysozyme (ground-grown)	1145	24111	0.242	1.048	0.184	1.33	0.11	0.15		Vaney *et al.* (1996)
			0.218		0.226		0.12			
HEW lysozyme (space-grown)	1141	21542	0.259	1.040	0.183	1.40	0.12			Vaney *et al.* (1996)
			0.230		0.226		0.13			
γB-crystallin	1708	26151	0.297	1.032	0.180	1.49	0.14	0.16	0.12	Tickle *et al.* (1998*a*)
			0.256		0.204		0.14			
βB2-crystallin	1558	18583	0.356	~1.032	0.184	2.10	0.25	0.21	0.17	Tickle *et al.* (1998*a*)
			0.290		0.200		0.22			
β-purothionin	439	4966	0.370	1.050	0.198	1.70	0.22	0.22		Stec, Rao *et al.* (1995)
			0.297		0.281		0.26			
α_1-purothionin	434	1168	†	1.180	0.155	2.50	—	0.25		Rao *et al.* (1995)
			0.610		0.218		0.68			
EM lysozyme	1068	8308	0.514	1.040	0.169	1.90	0.30	0.20	0.18	Guss *et al.* (1997)
			0.359		0.229		0.28			
Azurin II	1012	12162	0.353	1.174	0.188	1.90	0.26	0.15	0.25	Dodd *et al.* (1995)
			0.288		0.207		0.23			
Ribonuclease A with RI	4416	18859	1.922	1.145	0.194	2.50	1.85	0.32	0.57	Kobe & Deisenhofer (1995)
			0.484		0.286		0.69			
Fab HyHEL-5 with HEWL	4333	11754	†	1.111	0.196	2.65	—	0.30		Cohen *et al.* (1996)
			0.607		0.288		0.69			

† (N_i/p) negative.

As examples, the Luzzati plots for the two structures of TGF-β2 are shown in Fig. 8. Daopin *et al.* (1994) inferred average $\langle\Delta r\rangle$ of around 0.21 and 0.23 Å.

Of the three Luzzati assumptions summarized above, the most attractive is the third, which does not require the atoms to be identical nor the position errors to be small. For proteins, there are very obvious difficulties with assumption (ii). Errors do depend very strongly on Z and B. In the high-angle data shells, atoms with large B values contribute neither to ΔF nor to $|F|$, and so have no effect on R in these shells. In their important paper on protein accuracy, Chambers & Stroud (1979) wrote 'the [Luzzati] estimate derived from reflections in this range applies mainly to [the] best determined atoms'.

Thus, a Luzzati plot seems to allow a cautious upper limit statement about the precision of the best parts of a structure, but it gives little indication for the poor parts.

One reason for the popularity of Luzzati plots is that the R values for the middle and outer shells of a structure often roughly follow a Luzzati curve. Evidently the effective average $\langle\Delta r\rangle$ for the structure must be decreasing as $\sin\theta/\lambda$ increases since atoms of high B are ceasing to contribute, while the proportionate experimental and model errors must be increasing. This also suggests that the upper limit for $\langle\Delta r\rangle$ for the low-B atoms could be estimated from the lowest Luzzati theoretical curve touched by the experimental R plot. Thus, in Fig. 8 the upper limits for the low-B atoms could be taken as 0.18 and 0.21 Å, rather than the 0.21 and 0.23 Å chosen by Daopin *et al.* (1994).

Since the introduction of R_{free} by Brünger (1992) and the discussion of R_{free} by Tickle *et al.* (1998*b*), it can be seen that Luzzati plots should be based on a residual more akin to R_{free} rather than R in order to avoid bias from the fitting of data.

The mean positional error $\langle\Delta r\rangle$ of atoms can also be estimated from the σ_A plots of Read (1986, 1990). This method arose from Read's analysis of improved Fourier coefficients for maps using phases from partial structures with errors. It is preferable in several respects to the Luzzati method, but like Luzzati it assumes that the coordinate distribution is the same for all atoms. Luzzati and/or Read estimates of $\langle\Delta r\rangle$ are available for some of the structures in Tables 2 and 3. Often the two estimates are not greatly different.

8.2. Statistical reinterpretation of Luzzati plots

Luzzati plots are fundamentally different from other statistical estimates of error. The Luzzati theory applies to an idealized incomplete refinement and estimates the average shifts needed to reach $R = 0$. In the least-squares method the

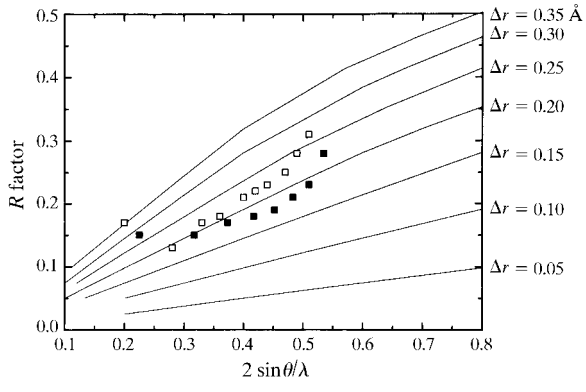

Figure 8
Luzzati plots showing the refined R factor as a function of resolution for 1TGI (solid squares) and 1TGF (open squares) from Daopin *et al.* (1994).

equations for shifts are quite different from the equations to estimate variances in completed refinements.

However, Luzzati-style plots of R *versus* $\sin\theta/\lambda$ can be reinterpreted to give statistically based estimates of $\sigma(x)$.

During Cruickshank's (1960) derivation of the approximate equation (23) for $\sigma(x)$ in diagonal least squares, he reached an intermediate equation

$$\sigma^2(x) = N_i/\left[4\sum_{\rm obs}(s^2/R^2)\right]. \quad (32)$$

He then assumed R to be independent of s, and took R outside the summation to obtain (23) above.

Luzzati (1952) calculated the acentric residual R as a function of s and $\langle\Delta r\rangle$, the average radial error of the atomic positions. His analysis shows that R is a linear function of s and $\langle\Delta r\rangle$ for a substantial range of $s\langle\Delta r\rangle$, with

$$R(s, \langle\Delta r\rangle) = (2\pi)^{1/2} s\langle\Delta r\rangle. \quad (33)$$

The theoretical Luzzati plots of R are nearly linear for small to medium $s = 2\sin\theta/\lambda$ (see Fig. 8). If we substitute this R in the least-squares estimate (32), a little manipulation along the lines of §6 (including the cautionary factor 1/0.65) then gives

$$\sigma_{\rm LS,Luzz}(x) = 1.93(N/p)^{1/2}\langle\Delta r\rangle_{\rm Luzz} \quad (34)$$

or

$$\sigma_{\rm LS,Luzz}(r) = 3.34(N/p)^{1/2}\langle\Delta r\rangle_{\rm Luzz}, \quad (35)$$

where $\langle\Delta r\rangle_{\rm Luzz}$ is the radial error found from the Luzzati plot and $\sigma_{\rm LS,Luzz}(r)$ is the least-squares estimate, assuming the final residual behaves as a function of s in the manner described by Luzzati in (33). Subject to the stated approximations, (35) is a proper statistical estimate of the diffraction-data-only s.u. $\sigma(r)$. As expected statistically, the number of observations, the number of parameters and the proportionate scattering power of a single atom enter into the result. These terms are absent from Luzzati's estimate of $\langle\Delta r\rangle_{\rm Luzz}$ from $R(s)$.

Authors using Luzzati plots, as in Fig. 8, usually follow Luzzati and give values of $\langle\Delta r\rangle_{\rm Luzz}$, the *average* positional error. The *r.m.s.* positional error (Luzzati, 1952) in a three-dimensional Gaussian distribution is $\sigma_{\rm Luzz}(r) = (3\pi/8)^{1/2}\langle\Delta r\rangle_{\rm Luzz} = 1.085\langle\Delta r\rangle_{\rm Luzz}$. From this and (35), we see that $\sigma_{\rm Luzz}(r) = \sigma_{\rm LS,Luzz}(r)$ when $(N/p)^{1/2} = 1.085/3.34 = 0.32$.

We can now see why reported Luzzati values of $\langle\Delta r\rangle$ were often plausible. For many proteins in Tables 2 and 3 $(N/p)^{1/2}$ is around 0.35, and $\sigma_{\rm Luzz}(r)$ and $\langle\Delta r\rangle_{\rm Luzz}$ are then similar in value to $\sigma_{\rm LS,Luzz}(r)$. For low-resolution studies with large $(N/p)^{1/2}$, the Luzzati value for a low-B atom is smaller than the DPI diffraction-only error for a $B_{\rm avg}$ atom. Conversely, for atomic resolution studies with $(N/p)^{1/2} < 0.2$, the Luzzati plot overestimates the diffraction errors [as Luzzati had supposed in 1952 (Luzzati, 1952)].

Using (33), equation (35) can be written

$$\sigma_{\rm LS,Luzz}(r) = 1.33(N/p)^{1/2}[R(s_m)/s_m], \quad (36)$$

where $R(s_m)$ is the value of R at some value of $s = s_m$ on the selected Luzzati curve.

Equation (36) provides a means of making a statistical estimate of error for an atom with $B = B_{\rm avg}$ (the average B for fully occupied sites) from a plot of R *versus* $2\sin\theta/\lambda$.

If the original published Luzzati values of $\langle\Delta r\rangle_{\rm Luzz}$ are substituted in (35), the resulting $\sigma_{\rm LS,Luzz}(r)$ are typically from 10% lower to 35% higher than the DPI $\sigma(r,B_{\rm avg})$ of Tables 2 and 3. Equality of these $\sigma(r)$ requires by (30) and (36) that

$$3^{1/2}Rd_{\rm min} = 1.33[R(s_{\rm max})/s_{\rm max}], \quad (37)$$

i.e. $R(s_{\rm max}) = 1.30\,R$. Here, $R(s_{\rm max})$ is not the actual value of R in the outermost shell, but is the value corresponding to the selected Luzzati curve. While one expects $R(s_{\rm max})$ to be about this much larger (or more) than a conventional averaged R, the implied value of $R(s_{\rm max})$ will have depended on where the original authors placed their Luzzati $\langle\Delta r\rangle$ curve among the scattered experimental points. Both (30) and (36) are based on gross simplifications, the one that $R(s)$ is constant and the other that $R(s)$ is proportional to s, so it is not surprising if the results, though agreeing in order of magnitudes, differ somewhat.

9. Discussion

9.1. Successive relaxation of restrictions

It has just been remarked in §8.2 that for large $(N/p)^{1/2}$ the Luzzati method yields a smaller error than the statistically derived diffraction-only DPI. Is this reasonable?

To compare the Luzzati method, the DPI and full-matrix inversion, we consider a series of problems in which constraints are successively relaxed. We note first that the Luzzati method can be regarded alternatively, not as indicating how far an idealized refinement still has to go, but as a providing an estimate of the differences between two models.

As a thought experiment, consider two proteins P and Q whose molecular structures are believed to be very similar, though not identical. Suppose P is a known structure, but Q is an unknown structure for which low-resolution diffraction data have been measured.

We can use P as a rigid-body model in the molecular-replacement method to find the position of Q within its crystal unit cell. Ignoring all difficulties about B values or disorder, we have a six-parameter problem with three coordinates for the translation and three for the rotation of the model P. The estimated statistical precision of the six parameters will be given by the inversion of a 6×6 least-squares matrix. The coordinate precision of individual atoms will be given by the functional dependence of these coordinates on the translation and rotation parameters. There is no full matrix of order $3n_{\rm atoms}$ in this problem.

However, we can produce a Luzzati plot from the observed and calculated structure factors for Q. This will give a good estimate of the average differences $\langle\Delta r\rangle$ between the P and Q molecules, provided the Luzzati assumptions are valid and the precisions of the six LS parameters are well below $\langle\Delta r\rangle$. In this thought experiment the approximate diffraction-only formula (30), with large N_i and rather few observations, could well

yield a $\sigma(r,B_{avg})$ much above the Luzzati $\langle\Delta r\rangle$. The fault is with the DPI (30).

A next stage of relaxation from the rigid-body model is to assume that the distances in the residues and side chains of Q are all fixed, *i.e. constrained* at the distances of protein P, but that the torsion angles of the main and side chains have to be determined. At the end of the refinement, the uncertainties in the molecular position and torsion angles of Q will be given by the inversion of a matrix of order $6 + n_{tors}$, where n_{tors} is the number of torsion angles. Individual coordinate uncertainties will be given through the complicated functional dependence of the coordinates on the torsion angles and molecular position.

Again, if there are relatively few observations, it would be possible for the DPI (30) to yield a $\sigma(r,B_{avg})$ above the Luzzati $\langle\Delta r\rangle$. However, the precision estimated from the matrix of order $6 + n_{tors}$ is formally correct rather than the Luzzati value.

The next stage of relaxation is to assume that the distances in the residues and side chains are *restrained* by weights such as those in the Engh & Huber (1991) scheme. At the end of the refinement, reached by whatever path, we can calculate a full matrix of order $4n_{atoms}$ built up from diffraction and restraint terms, which directly involves the individual x_i and B_i. Inversion of this matrix gives the precision of the coordinates. Clearly, the smaller $(N/p)^{1/2}$, *i.e.* the more observations per parameter, the poorer the Luzzati estimate will be relative to the individual $\sigma_{res}(r)$ from the full-matrix inversion.

It is a plausible conjecture for real problems that the Luzzati estimate of $1.085\langle\Delta r\rangle$ for low-B atoms will always be larger than the $\sigma(r_i)$ obtained from the inversion of the restrained full matrix using ΔF_h as an estimate of $\sigma(F_h)$.

9.2. α_1- and β-purothionin

Table 3 includes results for the 45-residue proteins α_1-purothionin (α-PT; Rao *et al.*, 1995) and β-purothionin (β-PT; Stec, Rao *et al.*, 1995) determined at 2.5 and 1.7 Å, respectively. From Luzzati plots, the Luzzati $\langle\Delta r\rangle$ are reported as 0.25 and 0.22 Å, respectively. Superficially, it is surprising that the low-resolution α_1-PT with 1168 reflections should seem almost as precise as the higher resolution β-PT with 4966 reflections. Each structure required nearly the same number of parameters, approximately $435 \times 4 = 1740$.

For α_1-PT, the number of reflections is less than the number of parameters, but the structure was solved successfully by including restraints on the bond and angle distances. No estimate of $\sigma(r,B_{avg})$ using R in the DPI (30) is possible because $p = n_{obs} - n_{params}$ is negative. However, (31) using R_{free} with $(N/n_{obs})^{1/2} = 0.610$ gives $\sigma(r,B_{avg}) = 0.68$ Å. This is well above the Luzzati 0.25 Å (which, as usual, relates to the best atoms in the structure rather than an average). For reasons discussed in §10.2, the true restrained $\sigma_{res}(r,B_{avg})$ is probably considerably better than 0.68 Å.

For β-PT, (30) using R with $(N/p)^{1/2} = 0.370$ gives 0.22 Å, while (31) using R_{free} with $(N/p)^{1/2} = 0.297$ gives 0.26 Å. These are close to the Luzzati $\langle\Delta r\rangle$ of 0.22 Å, which must be multiplied by 1.085 to convert to an r.m.s. value.

Note that in the case of β-PT, the published Luzzati plot gives much the same $\langle\Delta r\rangle$ whether the plot is terminated at the actual resolution limit of 1.7 Å or at the 2.5 Å limit of β-PT. The DPI surely correctly reflects an intuitive feeling that a quadrupling of the number of reflections must improve the precision of the results.

Full-matrix error estimates for these two structures would be very instructive.

9.3. More about restraints

Geometric restraint dictionaries typically use bond-length weights based on $\sigma_{geom}(l)$ of around 0.02 or 0.03 Å. Tables 1, 2 and 3 show that even 1.5 Å studies have diffraction-only errors $\sigma_{diff}(x,B_{avg})$ 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, as indicated earlier, the dictionary offers no values for many of the configurational parameters of the protein structure, including the centroid and molecular orientation.

If the protein main chain were represented by a chain of rigid peptide groups, 12 coordinate parameters per successive group would be reduced to two torsion angles per group. Off-diagonal terms between these torsion variables would be relatively weaker than those between restrained pairs of atoms. Even if successive peptide groups were treated as not coupled at the Cα atoms, each group could be specified by three coordinates and three orientation angles. Thus, one may suspect that in 1.5 Å and lower resolution restrained refinements the true atomic $\sigma(x)$ in peptides may be between $(2/12)^{1/2} = 0.4$ and $(6/12)^{1/2} = 0.7$ times the values given by the unrestrained full matrix. Indeed, it was noted in §4.1 for concanavalin A that for high B, where only low-resolution terms matter directly, the restrained $\sigma(r)$ were less than half the unrestrained values. Similar arguments apply to side chains.

Conversely, because of the existence of diffraction data to 0.94 Å resolution for concanavalin A, the precisions of the positions and bond lengths for atoms with low B were influenced more by the diffraction data than by the weights assigned to the stereochemical dictionary. Indeed, it was this high resolution of the data which made it possible (Deacon *et al.*, 1997) to distinguish in the unrestrained refinement between bond lengths 1.170 (9) Å and 1.324 (10) Å as C═O and C—O(H) in a key carboxyl group, Asp28. The result was confirmed by the detection of the H atom both in the X-ray and in a neutron Laue analysis (Habash *et al.*, 1997).

10. Final remarks

10.1. Luzzati plots

Protein structures always show a great range of B values. The Luzzati theory effectively assumes that all atoms have the same B. Nonetheless, the Luzzati method applied to high-angle data shells does provide an upper limit for $\langle\Delta r\rangle$ for the atoms with low B. It is an upper limit since experimental errors and model imperfections are not allowed for in the theory.

Low-resolution structures can validly be determined by using restraints, even though the number of diffraction observations is fewer than the number of atomic coordinates. The Luzzati method, based preferably on R_{free}, can be applied to the low-B atoms in such structures. As the number of observations increases and the resolution improves, the Luzzati $\langle\Delta r\rangle$ increasingly overestimates the true $\sigma(r)$ of the low-B atoms.

In the use of Luzzati plots, the method of refinement and its degree of convergence are irrelevant. A Luzzati plot is a statement for the low-B atoms about the maximum errors associated with a given structure, whether converged or not.

10.2. The diffraction-component precision index

The DPI, (30) or (31), provides a very simple formula for $\sigma(r,B_{avg})$. It is based on a very rough approximation to a diagonal element of the diffraction-data-only matrix. Using a diagonal element is a reasonable approximation for atomic resolution structures, but for low-resolution structures there will be significant off-diagonal terms between overlapping atoms. The effect can be simulated in the two-atom protein model of §3.2 by introducing positive off-diagonal elements into the diffraction-data matrix (6). As expected, $\sigma^2_{diff}(x_i)$ is increased. Therefore, the DPI will be an underestimate of the diffraction component in low-resolution structures.

However, the true restrained variance $\sigma^2_{res}(x_i)$ in the new counterpart of (15) remains less than the diagonal diffraction result (14) $\sigma^2_{diff}(x_i) = 1/a$. Thus, for low-resolution structures the DPI should be an overestimate of the true precision given by a restrained full-matrix calculation (where the restraints act to hold the overlapping atoms apart). This is confirmed by the results for the 2.1 Å study of β-B2-crystallin (Tickle *et al.*, 1998*a*) discussed in §7.3 and Table 3. The restrained full-matrix average for all protein atoms was $\sigma_{res}(r) = 0.15$ Å, compared with the DPI of 0.25 Å (on R) or 0.22 Å (on R_{free}). The ratio between the unrestrained DPI and the restrained full-matrix average is consistent with the discussion in §9.3 of a low-resolution protein as a chain of effectively rigid peptide groups. The ratio no doubt becomes much worse for resolutions of 3 Å and above.

The DPI estimate of $\sigma(r,B_{avg})$ is given by a formula of 'back-of-an-envelope' simplicity. B_{avg} is taken to be the average B for fully occupied sites, but the weights implicit in the averaging are not well defined in the derivation of the DPI. Thus, the DPI should perhaps be regarded as simply offering an estimate of a typical $\sigma_{diff}(r)$ for a C or N atom with a mid-range B. From the evidence of Tables 1, 2 and 3, except at low resolution it seems to give a useful overall indication of protein precision even in restrained refinements.

The DPI evidently provides a method for the comparative ranking of different structure determinations. In this regard it is a complement to the general use of d_{min} as a quick indicator of possible structural quality.

Note that (24) and (25) offer scope for making individual error estimates for atoms of different B and Z.

10.3. Restrained refinement

Compared with unrestrained refinements for small molecules, in restrained refinements for proteins there is a major numerical distinction between the s.u. $\sigma(x_i)$ of an atomic coordinate and the s.u.$\sigma(l_{ij})$ of a bond length. The atomic $\sigma(x_i)$ depends strongly on B or **B** and may be anisotropic (owing to **B** or the geometry of the restraints). However, except in rather high-resolution work such as concanavalin A (Fig. 4*a*), $\sigma(l_{ij})$ is not far from the $\sigma_{geom}(l)$ of the restraint weighting. Small r.m.s. differences between refined and dictionary bond lengths are not an indication of $\sigma(x)$ quality. [See Tickle *et al.* (1998*a*,*b*) for an algebraic analysis of these differences.]

Reliable $\sigma(x)$ values are needed for any discussion of non-dictionary distances between atoms in different residues, between protein and solvent atoms or between metal atoms and their ligands.

10.4. Fourier map formula

The Fourier map formula (1) has the great advantage over Luzzati plots and the DPI that it provides directly a strong dependence of $\sigma(x)$ on B. However, it requires summations not customarily provided in protein program suites. Similar precision estimates may, however, be obtained from the diagonal elements of the diffraction-data-only least-squares matrix.

Despite the good results obtained from (1) by Daopin *et al.* (1994), such estimates lack consideration of overlap and restraint effects. Rather than seek rough approximations to deal with these, it is more sensible to calculate block matrices.

10.5. Full-matrix estimates of precision

The original contention of this paper in §1 was that the variances and covariances of the structural parameters of proteins can be found from the inverse of the least-squares normal matrix. However, there was a caveat (§1), stating chiefly that explicit account would not be taken of disorder of the solvent or of parts of the protein. Correction by Babinet's principle of complementarity is only a crude first-order approximation. The consequences of such disorder problems, which make the variation of calculated structure factors non-linear over the range of interest, may in future be better handled by maximum-likelihood methods (*e.g.* Bricogne, 1993; Bricogne & Irwin, 1996; Murshudov *et al.*, 1997; Read, 1990). 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 paper are probably somewhat low.

As noted several times above, the formidable computational task of assembling and inverting the full matrix has already been accomplished in a number of analyses. With the increasing power of computers, a final diffraction-cum-restraints 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, two partial schemes can be suggested. As mentioned in §5, 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 comparable scheme in restrained refinements of any resolution might be 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.

APPENDIX *A*
Some least-squares and Fourier method formulae

A1. The least-squares method

In the unrestrained least-squares method, the residual

$$R = \sum_3 w(hkl)\Delta^2(hkl) \tag{38}$$

is minimized, where Δ is either $|F_o| - |F_c|$ for R_1 or $|F_o|^2 - |F_c|^2$ for R_2 and $w(hkl)$ is chosen appropriately. The summation is over independent planes. There are good reasons for preferring R_2, as is done in *SHELXL*. However, for notational simplicity R_1 will be used here. Also, (hkl) will be omitted.

When R is a minimum with respect to the parameter u_j, $\partial R/\partial u_j = 0$, *i.e.*

$$\sum_3 w\Delta(\partial\Delta/\partial u_j) = 0 \ \text{ or } \sum_3 w\Delta(\partial|F_c|/\partial u_j) = 0, \tag{39}$$

since $\partial\Delta/\partial u_j = -\partial|F_c|/\partial u_j$. For a trial set of parameters close to the correct set, the normal equations for the corrections ε_j to the n parameters u_j are the n simultaneous linear equations given by considering first-order changes in the Δ owing to the ε_j,

$$\sum_i \varepsilon_i \left(\sum_3 w \frac{\partial|F_c|}{\partial u_i}\frac{\partial|F_c|}{\partial u_j} \right) = \sum_3 w\Delta \frac{\partial|F_c|}{\partial u_j}. \tag{40}$$

This can be abbreviated to

$$\sum_i \varepsilon_i a_{ij} = b_j. \tag{41}$$

Some important points in the derivation of the s.u.s of the refined parameters can be most easily understood if we suppose that the matrix a_{ij} can be approximated by its diagonal elements. Each parameter is then determined by a single equation of the form

$$\varepsilon_i \sum_3 wg^2 = \sum_3 wg\Delta, \tag{42}$$

where $g = \partial|F_c|/\partial u_i$. Hence,

$$\varepsilon_i = \left(\sum_3 wg\Delta\right) \Big/ \left(\sum_3 wg^2\right), \tag{43}$$

so that the variance of the parameter is

$$\sigma_i^2 = \left[\sum_3 w^2 g^2 \sigma^2(F)\right] \Big/ \left(\sum_3 wg^2\right)^2. \tag{44}$$

If the weights have been chosen as $w(hkl) = 1/\sigma^2(F_{hkl})$, this simplifies to

$$\sigma_i^2 = 1 \Big/ \left(\sum_3 wg^2\right) = 1/a_{ii}, \tag{45}$$

which is appropriate for absolute weights. Equation (45) provides an s.u. for a parameter relative to the s.u.s $\sigma(F)$ of the observations.

In general, with the full matrix a_{ij} in the normal equations

$$\sigma_i^2 = (a^{-1})_{ii}, \tag{46}$$

where $(a^{-1})_{ii}$ is an element of the matrix inverse to a_{ij}. The covariance of the parameters u_i and u_j is $\mathrm{cov}(i,j) \equiv \sigma_i\sigma_j\mathrm{correl}(i,j) = (a^{-1})_{ij}$.

In the early stages of refinement, artificial weights may be chosen to accelerate refinement. In the final stages, the weights must be related to the precision of the structure factors if parameter variances are being sought. There are two distinct ways, covering two ranges of error, in which this may be performed.

(i) The weights may reflect the precision of the $|F_o|$, so that

$$w(hkl) = 1/\sigma^2(F_{hkl}) \tag{47}$$

where σ^2 is the estimated variance of $|F_o|$ owing to a specific class of random experimental errors. These absolute weights are determined from an analysis of the experiment. Weights chosen in this way lead to estimated parameter variances

$$\sigma_i^2 = (a^{-1})_{ii}, \tag{46}$$

which cover only the specific class of random experimental error.

(ii) The weights may reflect the trends in $|\Delta| \equiv \big||F_o| - |F_c|\big|$. A weighting function with a small number of parameters is chosen so that the averages of $w\Delta^2$ are constant when the set of $w\Delta^2$ values is analysed in any pertinent fashion (*e.g.* in bins of increasing $|F_o|$ and $\sin\theta/\lambda$). Weights chosen in this way are relative weights and the expression for the parameter variances needs a scaling factor

$$S^2 = \left(\sum_3 w\Delta^2\right)/(n_{\mathrm{obs}} - n_{\mathrm{params}}). \tag{48}$$

Hence, in the full-matrix case

$$\sigma_i^2 = \left[\left(\sum_3 \Delta^2\right)/(n_{\mathrm{obs}} - n_{\mathrm{params}})\right](a^{-1})_{ii}, \tag{49}$$

which allows for all random experimental errors, such systematic experimental errors as cannot be simulated in the $|F_c|$ and imperfections in the calculated model.

Cruickshank's (1960) order-of-magnitude formula for $\sigma(x)$ quoted above at equation (23) was derived from the diagonal

form of (47) by gross approximations to $\sum w\Delta^2$ and $a_{ii} \equiv \sum wg^2$.

A2. The modified Fourier method

Cochran (1948) showed that the coordinates (x_r, ...) of atom r which minimize

$$\varphi = \sum_3 (1/f_r)(|F_o| - |F_c|)^2 \quad (50)$$

are the same as those found from the Fourier series for the electron density,

$$\rho_0 = (1/V) \sum_{hkl} |F_o| \cos(\theta - \alpha), \quad (51)$$

when this is corrected for finite summation and peak overlapping by a ρ_c series. In (50), the scattering factor f_r includes the vibration exponential. In (51), $\theta = 2\pi(hx + \ldots)$. Note that in (50) the summation is over independent planes, whereas in (51) it is over independent planes and their symmetry equivalents. Anomalous scattering will be neglected in the following discussion.

The key to Cochran's result is that the condition $\partial\varphi/\partial x_r = 0$ yields the result

$$-2\pi \sum_{hkl} h\Delta \sin(\theta_r - \alpha) = 0. \quad (52)$$

This follows because $|F_c|$ involves atom r and its symmetry equivalents, so that differentiation with respect to x_r throws up terms like $-2\pi h f_r \sin(\theta r - \alpha)$ and symmetry equivalents. The latter can be reformulated as contributions from equivalent planes, thus changing the summation from $\sum_3$ to $\sum_{hkl}$. The f_r is cancelled by the artificial weight $1/f_r$ in φ. Details of the calculation will be found in Cruickshank (1952).

By differentiation of Fourier series such as (51), the condition (52) can then be interpreted as $[\partial(\rho_o - \rho_c)/\partial x]_r = 0$, *i.e.* the slope of the difference map at the position of atom r is zero. Equivalently, the slopes at atom r of the observed and calculated electron densities are equal. As a criterion this becomes the basis of the modified Fourier method (Cruickshank, 1952, 1959), which like the least-squares method is applicable whether the atomic peaks are resolved or not. 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.

For centrosymmetric structures, the diagonal term multiplying a correction ε_{xr} is

$$(\partial/\partial x_r)(\partial\rho_r/\partial x)_r, \quad (53)$$

where ρ_r is the contribution to ρ_c of atom r and its symmetry equivalents and the derivative is evaluated at the trial position of atom r. Evaluation of (53) yields a dominant term which is a second-derivative series proportional to

$$\sum_{hkl} h^2 f_r. \quad (54)$$

For acentric reflections, the dependence of the phase angle α on x_r also has to be considered. The statistical argument (Cruickshank, 1952) is a little tedious, but the outcome for non-centrosymmetric structures is that the diagonal term of the modified Fourier method is proportional to

$$\sum_{hkl} h^2 f_r(m/2), \quad (55)$$

where m = 1 or 2 for acentric or centric reflections and f_r here includes the vibration exponential. This is the origin of the atomic peak 'curvature' term (3) in the Fourier map approach to estimation of the error $\sigma(x)$ by (1).

The σ(slope) term (2) is simply an estimate of the r.m.s. error at a general position (Cruickshank & Rollett, 1953) in the slope of the difference map, *i.e.* the r.m.s. error on the right-hand side of the modified Fourier method.

There is a relation between minimization of $R_2 = \sum_3 w(|F_o|^2 - |F_c|^2)^2$ and refinement by a modified Patterson method (Cruickshank, 1952).

A3. Relative weighting of diffraction and restraint terms

When only relative diffraction weights are known, as in (49), it has been customary (Rollett, 1970) to scale the 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 $S^2 = (\sum_3 w_h \Delta_h^2)/(n_{\rm obs} - n_{\rm params})$. 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 restraint weights are on an absolute scale, they propose that the (absolute) scale of the diffraction weights should be determined by adjustment until the restrained residual R' (4) is equal to its expected value $(n_{\rm obs} + n_{\rm restraints} - n_{\rm params})$.

A4. Statistical descriptors and goodness of fit

In recent years, there have been developments and changes in statistical nomenclature and usage. Many aspects are summarized in the Reports of IUCr subcommittees on Statistical Descriptors in Crystallography (Schwarzenbach *et al.*, 1989, 1995). In their second Report, *inter alia* they emphasize the terms *uncertainty* and *standard uncertainty* (s.u.). The latter is a replacement for the older term *estimated standard deviation* (e.s.d.).

The expression S^2, (48) above, is called the *goodness of fit* when the weights are the reciprocals of the absolute variances of the observations.

One recommendation in the second Report does call for comment here. While agreeing that formulae such as (49) lead to conservative estimates of parameter variances, the Report suggests this practice is based on the questionable assumption that the variances of the observations by which the weights are assigned are relatively correct but uniformly underestimated. When the goodness of fit $S > 1$, then either the weights or the model or both are suspect.

I comment. My account in §A.1 describes two distinct ways of estimating parameter variances, covering two ranges of error. The kind of weights envisaged in the Report [based on

variances of Type A (estimated statistically) and of Type B (estimated otherwise)] are of a class described for Method I. They are not the weights to be used in Method II. Method II implicitly assumes from the outset that there are experimental errors, some covered and others not covered by Method I, and that there are imperfections in the calculated model. It avoids exploring the relative proportions and details of these sources, and aims to provide a realistic estimate of parameter uncertainties which can be used in external comparisons. It can be formally objected that Method II does not conform to the criteria of random variable theory, since clearly the Δs are partially correlated through the remaining model errors and some systematic experimental errors. However, it is a useful procedure. Method I on its own would present an optimistic view of the reliability of the overall investigation, the degree of optimism being indicated by the inverse of the goodness of fit (48). In Method II, if the weights are on an arbitrary scale then S^2 can have an arbitrary value.

For an advanced-level treatment of many aspects of the Refinement of Structural Parameters see §8 of *International Tables for Crystallography*, *Volume C* (1992). §8.5 is on the detection and treatment of systematic error.

I am especially grateful to G. M. Sheldrick, J. R. Helliwell and A. Deacon for the full-matrix calculations described in §4 and for their comments. I also acknowledge very useful comments on earlier drafts by D. M. Blow, E. J. Dodson, H. C. Freeman, M. M. Harding, W. N. Hunter, L. H. Jensen, D. S. Moss, P. Daopin Sun and M. R. Truter.

References

Allen, F. H., Cole, J. C. & Howard, J. A. K. (1995*a*). *Acta Cryst.* A**51**, 95–111.

Allen, F. H., Cole, J. C. & Howard, J. A. K. (1995*b*). *Acta Cryst.* A**51**, 112–121.

Blackwell, K. A., Anderson, B. F. & Baker, E. N. (1994). *Acta Cryst.* D**50**, 263–270.

Bode, W. & Schwager, P. (1975). *J. Mol. Biol.* **98**, 693–717.

Bricogne, G. (1993). *Acta Cryst.* D**49**, 37–60.

Bricogne, G. & Irwin, J. (1996). *Macromolecular Refinement. Proceedings of the CCP4 Study Weekend*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.

Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.

Chambers, J. L. & Stroud, R. M. (1979). *Acta Cryst.* B**35**, 1861–1874.

Cochran, W. (1948). *Acta Cryst.* **1**, 138–142.

Cohen, G. H., Sheriff, S. & Davies, D. R. (1996). *Acta Cryst.* D**52**, 315–326.

Cox, E. G. & Cruickshank, D. W. J. (1948). *Acta Cryst.* **1**, 92–93.

Cruickshank, D. W. J. (1949*a*). *Acta Cryst.* **2**, 65–82.

Cruickshank, D. W. J. (1949*b*). *Acta Cryst.* **2**, 154–157.

Cruickshank, D. W. J. (1952). *Acta Cryst.* **5**, 511–518.

Cruickshank, D. W. J. (1956). *Acta Cryst.* **9**, 747–754.

Cruickshank, D. W. J. (1959). *International Tables for X-ray Crystallography*, Vol. 2, edited by J. S. Kasper & K. Lonsdale, pp. 318–340. Birmingham: Kynoch Press.

Cruickshank, D. W. J. (1960). *Acta Cryst.* **13**, 774–777.

Cruickshank, D. W. J. (1965). *Acta Cryst.* **19**, 153.

Cruickshank, D. W. J. (1996*a*). *Macromolecular Refinement. Proceedings of the CCP4 Study Weekend*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 11–22. Warrington: Daresbury Laboratory.

Cruickshank, D. W. J. (1996*b*). *Acta Cryst.* A**52**, C85.

Cruickshank, D. W. J. & Robertson, A. P. (1953). *Acta Cryst.* **6**, 698–705.

Cruickshank, D. W. J. & Rollett, J. S. (1953). *Acta Cryst.* **6**, 705–707.

Daopin, S., Davies, D. R., Schlunegger, M. P. & Grütter, M. G. (1994). *Acta Cryst.* D**50**, 85–92.

Dauter, Z., Sieker, L. C. & Wilson, K. S. (1992). *Acta Cryst.* B**48**, 42–59.

Deacon, A., Gleichmann, T., Kalb (Gilboa), A. J., Price, H., Raftery, J., Bradbrook, G., Yariv, J. & Helliwell, J. R. (1997). *J. Chem. Soc. Faraday Trans.* **93**, 4305–4312.

Dodd, F. E., Hasnain, S. S., Abraham, Z. H. L., Eady, R. R. & Smith, B. E. (1995). *Acta Cryst.* D**51**, 1052–1064.

Dodson, E., Kleywegt, G. J. & Wilson, K. (1996). *Acta Cryst.* D**52**, 228–234.

Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.

Fields, B. A., Bartsch, H. H., Bartunik, H. D., Cordes, F., Guss, J. M. & Freeman, H.C. (1994). *Acta Cryst.* D**50**, 709–730.

Frazão, C., Soares, C. M., Carrondo, M. A., Pohl, E., Dauter, Z., Wilson, K. S., Hervás, M., Navarro, J. A., De La Rosa, M. A. & Sheldrick, G. M. (1995). *Structure*, **3**, 1159–1169.

Guss, J. M., Bartunik, H. D. & Freeman, H. C. (1992). *Acta Cryst.* B**48**, 790–811.

Guss, J. M., Messer, M., Costello, M., Hardy, K. & Kumar, V. (1997). *Acta Cryst.* D**53**, 355–363.

Habash, J., Raftery, J., Weisgerber, S., Cassetta, A., Lehmann, M. S., Høghøj, P., Wilkinson, C., Campbell, J. W. & Helliwell, J. R. (1997). *J. Chem. Soc. Faraday Trans.* **93**, 4313–4317.

Haridas, M., Anderson, B. F. & Baker, E. N. (1995). *Acta Cryst.* D**51**, 629–646.

Hendrickson, W. A. & Konnert, J. H. (1980). *Computing in Crystallography*, edited by R. Diamond, S. Ramaseshan & K. Venkatesan, pp. 13.01–13.23. Bangalore: Indian Academy of Sciences.

International Tables for Crystallography, Volume C (1992). Dordrecht: Kluwer Academic Publishers.

Ko, T.-P., Day, J., Greenwood, A. & McPherson, A. (1994). *Acta Cryst.* D**50**, 813–825.

Kobe, B. & Deisenhofer, J. (1995). *Nature (London)*, **374**, 183–186.

Langridge, R., Marvin, D. A., Seeds, W. E., Wilson, H. R., Hooper, C. W., Wilkins, M. H. F. & Hamilton, L. D. (1960). *J. Mol. Biol.* **2**, 38–64.

Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.

Peters-Libeu, C. & Adman, E. T. (1997). *Acta Cryst.* D**53**, 56–77.

Rao, U., Stec, B. & Teeter, M. M. (1995). *Acta Cryst.* D**51**, 904–913.

Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.

Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.

Rollett, J. S. (1970). *Crystallographic Computing*, edited by F. R. Ahmed, S. R. Hall & C. P. Huber, pp. 167–181. Copenhagen: Munksgaard.

Schwarzenbach, D., Abrahams, S. C., Flack, H. D., Gonschorek, W., Hahn, T., Huml, K., Marsh, R. E., Prince, E., Robertson, B. E., Rollett, J. S. & Wilson, A. J. C. (1989). *Acta Cryst.* A**45**, 63–75.

Schwarzenbach, D., Abrahams, S. C., Flack, H. D., Prince, E. & Wilson, A. J. C. (1995). *Acta Cryst.* A**51**, 565–569.

Sevcik, J., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1996). *Acta Cryst.* D**52**, 327–344.

Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319–343.

Stec, B., Rao, U. & Teeter, M. M. (1995). *Acta Cryst.* D**51**, 914–924.
Stec, B., Zhou, R. & Teeter, M. M. (1995). *Acta Cryst.* D**51**, 663–681.
Stroud, R. M. & Fauman, E. B. (1995). *Protein Sci.* **4**, 2392–2404.
Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998*a*). *Acta Cryst.* D**54**, 243–252.
Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998*b*). *Acta Cryst.* D**54**, 547–557.
Trueblood, K. N., Bürgi, H.-B., Burzlaff, H., Dunitz, J. D., Gramaccioli, C. M., Schulz, H. H., Shmueli, U. & Abrahams, S. C. (1996). *Acta Cryst.* A**52**, 770–781.
Usón, I., Pohl, E., Schneider, T. R., Dauter, Z., Schmidt, A., Fritz, H.-J. & Sheldrick, G. M. (1999). In preparation.
Vaney, M. C., Maignan, S., Riès-Kautt, M. & Ducruix, A. (1996). *Acta Cryst.* D**52**, 505–517.
Watenpaugh, K. D., Sieker, L. C., Heriott, J. R. & Jensen, L. H. (1973). *Acta Cryst.* B**29**, 943–956.
Wilson, A. J. C. (1950). *Acta Cryst.* **3**, 397–398.