Acta Cryst. (1998). D**54**, 243–252

Error Estimates of Protein Structure Coordinates and Deviations from Standard Geometry by Full-Matrix Refinement of γB- and βB2-Crystallin

Ian J. Tickle, Roman A. Laskowski and David S. Moss*

Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, England.
E-mail: d.moss@mail.cryst.bbk.ac.uk

(*Received 17 March 1997; accepted 22 July 1997*)

Abstract

Faster workstations with larger memories are making error estimation from full-matrix least-squares refinement a more practicable technique in protein crystallography. Using minimum variance weighting, estimated standard deviations of atomic positions have been calculated for two eye lens proteins from the inverse of a least-squares normal matrix which was full with respect to the coordinate parameters. γB-crystallin, refined at 1.49 Å yielded average errors in atomic positions which ranged from 0.05 Å for main-chain atoms to 0.27 Å for unrestrained water molecules. The second structure used in this work was that of βB2-crystallin refined at 2.1 Å resolution where the corresponding average errors were 0.08 and 0.35 Å, respectively. The relative errors in atomic positions are dependent on the number and kinds of restraints used in the refinements. It is also shown that minimum variance weighting leads to mean-square deviations from target geometry in the refined structures which are smaller than the variances used in the distance weighting.

1. Introduction

The random errors associated with the parameters of a refined crystal structure can be expressed in terms of estimated standard deviations (e.s.d.'s). These can be calculated and quoted for the x, y and z coordinates and root-mean-square displacements (U values) of each atom in the model structure. In small-molecule refinement e.s.d.'s are routinely calculated and indeed are required by some journals as a precondition for publication. For large molecules such as proteins, however, the calculations have not been regularly performed to date, being considered too memory- and computationally intensive.

The earliest methods for estimating errors in atomic positions employed gradients in electron density difference maps (Cruickshank, 1949*a*,*b*) and such techniques have recently been revisited by Daopin *et al.* (1994). More often e.s.d.'s are obtained during structure refinement by calculating the least-squares covariance matrix from the inverse of the normal equations matrix (Cruickshank, 1965) as is implemented, for example, in the small-molecule refinement package *SHELX* (Sheldrick, 1976, 1985, 1986; Robinson & Sheldrick, 1988). For proteins the matrices involved are very large (typically several thousand rows and columns), and the inversion, which is an order n^3 process, soon becomes unfeasible. Over the past ten years protein crystallographers have used various techniques for estimating the random errors in the coordinates of their model structures. These include the Luzzati plot (Luzzati, 1952) and the σ_A plot of Read (1986) which both provide an estimate of the average positional error, $|\overline{\Delta\mathbf{r}}|$, of a structure's coordinates. The Luzzati method assumes that the positional errors are normally distributed and that they alone account for the differences between the $|F_{\rm obs}|_r$ and $|F_{\rm calc}|_r$ for all reflections (Cruickshank, 1996). Theoretical relationships between scattering angle and R factor, assuming different values of $|\overline{\Delta\mathbf{r}}|$, are then used to estimate the average error from the observed data.

More recent attempts to estimate errors have included: the 'residue R factor' of Jones *et al.* (1991); the tabulated R indices of Elango & Parthasarathy (1990); the refinement protocol of Carson *et al.* (1994) which uses temperature factors, real-space fit residuals, geometric strains, dihedral angles and shifts from the previous refinement cycle; the 'discriminator' of Sevcik *et al.* (1993) which assesses the likely errors on each atom in terms of its temperature factor divided by its electron density in the final $2|F_o| - |F_c|$ map, or U/ρ where ρ is the electron density; an empirically derived six-parameter equation of Stroud & Fauman (1995), and the use of the diagonal elements of the inverse normal matrix in a final cycle of unrestrained least-squares refinement to give an estimate of the radial errors in atomic positions (Holland *et al.*, 1990).

Various studies have been performed to assess the accuracy of these estimation procedures. Fields *et al.* (1994) performed two independent refinements, using different refinement programs, using a single set of synchrotron X-ray data at 1.6 Å resolution. The r.m.s. differences between the final models were 0.08 Å for all C$^\alpha$ atoms, 0.08 Å for all backbone atoms and 0.12 Å for all non-H atoms (excluding six obvious outliers). The estimated maximum average error from the Luzzati plot was found to be 0.13 Å. Daopin *et al.* (1994) compared four different estimation methods, two for calculating local errors and two (the Luzzati and σ_A plots) for overall errors, finding the methods to be in good agreement. On the other hand, Ohlendorf (1994) compared four independently refined X-ray crystal structures of human interleukin 1β, first re-refining them against a common data set to minimize the effects of different data sets and refinement protocols. He found that the final structures differed from one another by 0.84 Å, which was roughly three times the error predicted by the Luzzati plots. Murshudov & Dodson (1997) have developed the theory of Cruickshank (1996) and used it to explore the relationship between temperature factors and positional errors estimated from a diagonal second derivative matrix.

In this paper, we have calculated e.s.d.'s of positional parameters for two proteins, using the full matrix of the normal equations of least-squares refinement. We will describe how these calculations have been implemented in a least-squares refinement program and will present the results for two trial proteins: γB- and βB2-crystallin. In particular, we analyse the relationship between atomic U values and the positional e.s.d.'s.

1.1. *Theory of least squares*

In order to understand the error analysis in the present work, it is necessary to fully comprehend the statistical basis of least-squares refinement, its fundamental assumptions and the consequences when these assumptions fail to hold. The proofs of most of the basic properties of least squares discussed below may be found in Hamilton (1964).

Classical least squares assumes that we are given unbiased observations drawn from a population whose errors have finite second moments. It further assumes that there is a known linear relationship between the observations and the parameters to be determined. For a unique least-squares estimate of these parameters, the rank of the matrix expressing this relationship must be at least equal to the number of parameters to be determined.

Minimization of a quadratic form weighted with the variance–covariance matrix of the observations then yields unbiased minimum variance parameter estimates. A special case of least squares occurs where the observations have errors which are normally distributed. In this case, the parameter estimates are also maximum-likelihood estimates. However, unlike the maximum-likelihood method, the least-squares method does not require any distributional assumptions regarding the observation errors other than the existence and knowledge of the second moments of the error distribution.

1.2. *Limitations of least squares in macromolecular refinement*

We now review these assumptions with respect to macromolecular crystallography and their relevance to the current work. First the structure-factor model expresses a non-linear relationship between the observations and the molecular parameters. This limits least-squares refinement to a refinement role rather than of structure determination. A second assumption concerns the accuracy of the model. Two types of inaccuracy may be identified. First, the structure factors may be calculated on the basis of an incomplete model where atoms are missing. In this work ordered solvent atoms and all protein atoms (apart from a few seriously disordered residues near the N and C termini in γB-crystallin) are present in the model. The contribution of disordered solvent to the structure amplitudes has been modelled by following the principle of Babinet (1837) and modelling the solvent continuum by dummy atoms. These are sited at protein atomic positions and have large temperature factors (Driessen *et al.*, 1989). A second type of model error concerns the functional form of the model. Disorder has been modelled in terms of isotropic harmonic displacement and this is certainly a poor model for the more mobile side chains where atomic positions might be more realistically represented by possibly anharmonic and multimodal distributions.

Errors in the functional form can be classified according to whether or not they can be simulated by parameter adjustment. Errors that can be simulated by parameter adjustment may be called 'silent errors' because they will make little contribution to the residuals which form the basis of agreement statistics such as R factors or correlation coefficients. Isotropic absorption may be such an error. Failure to correct either the model or the observations for absorption effects will lead to underestimated temperature factors but will not add much to the R factor.

Other model errors, such as departures from the harmonic model of atomic displacements, will produce effects which are nearly equivalent to introducing random errors into the observed data. These errors cannot be well simulated by parameter adjustment and are probably the cause of the higher R factors observed in macromolecular crystallography as compared with the crystallography of well ordered inorganic structures. Such errors are certainly significant in the present work and are treated as equivalent to errors in the observed data.

The presence of the latter errors means that structure-amplitude weighting cannot be based on experimental standard deviations but must reflect model errors. The determination of these weights is described in the next section. Incorrect weighting still results in unbiased parameter estimates but these estimates have larger variances than those calculated with correctly weighted observations.

2. Methods

2.1. *Calculation of e.s.d's*

The e.s.d.'s of the atomic coordinates were calculated from the inverse matrix $\mathbf{H}^{-1}$ obtained from the normal equations formed during the final cycle of least-squares refinement.

The least-squares refinement program *RESTRAIN* (Moss & Morffew, 1982; Haneef *et al.*, 1985; Driessen *et al.*, 1989) was modified to compute the full normal equations matrix **H** with respect to the positional parameters. Temperature factors were treated with the diagonal approximation.

The function minimized was

$$M = \sum_{r}^{N_{\rm refl}} w_r(|F_{\rm obs}|_r - G|F_{\rm calc}|_r)^2 + \sum_{s}^{N_{\rm dist}} w_s[d_{\rm target}(s) - d_{\rm calc}(s)]^2 + \sum_{t}^{N_{\rm planes}} w_t e_t. \quad (1)$$

The factor G is a scale factor between the two sets of structure-factor amplitudes. The weights w_r are the structure-amplitude weights whose relative scale was determined from an empirical structure-amplitude weighting scheme (Nielsen, 1977) whose parameters were adjusted to maximize the entropy

$$-\sum_{r}^{N_{\rm refl}} P_r \ln[P_r], \quad (2)$$

where

$$P_r = \frac{w_r(|F_{\rm obs}|_r - G|F_{\rm calc}|_r)^2}{\sum_{r}^{N_{\rm refl}} w_r(|F_{\rm obs}|_r - G|F_{\rm calc}|_r)^2}. \quad (3)$$

The absolute scale of the structure-amplitude weights was determined by adjustment until the residual M was equal to its expected value of $n - m$, where n is the number of observations (including restraints) and m is the number of parameters.

The $d_{\rm target}$ and $d_{\rm calc}$ are the target and calculated restraint distances, respectively. The weights w_s of the distance restraints were chosen as the reciprocals of the variances of the deviations from the target distances as obtained from small-molecule structures (Engh & Huber, 1991). The $d_{\rm target}$ distances themselves were obtained from the same source.

The planar restraint terms $w_t e_t$ are the weighted minimum eigenvalues of the product–moment matrices of the coordinates of the atoms involved in the planes. These eigenvalues are the sums of the squares of the planar deviations. Use of this term in the expression minimized, has the advantage that it does not restrain the plane to the plane determined in the previous iteration. No restraints were applied to the peptide planes. The function M was minimized by solving the normal equations

$$\mathbf{H}\Delta\mathbf{p} = \mathbf{v}, \quad (4)$$

where $\Delta\mathbf{p}$ is the vector of unknown Δp_i values. The elements of **H** and **v** are

$$H_{ij} = \sum_{r=1}^{N_{\rm refl}} w_r \frac{\partial|F|_r}{\partial p_i}\frac{\partial|F|_r}{\partial p_j} + \sum_{s=1}^{N_{\rm dist}} w_s \frac{\partial d_s}{\partial p_i}\frac{\partial d_s}{\partial p_j} + \sum_{t=1}^{N_{\rm planes}} w_t \frac{\partial^2 e_t}{\partial p_i \partial p_j} \quad (5)$$

$$v_i = \sum_{r=1}^{N_{\rm refl}} w_r(|F_{\rm obs}|_r - G|F_{\rm calc}|_r)\frac{\partial|F|_r}{\partial p_i} + \sum_{s=1}^{N_{\rm dist}} w_s[d_{\rm target}(s) - d_{\rm calc}(s)]\frac{\partial d_s}{\partial p_i} - \sum_{t=1}^{N_{\rm planes}} w_t \frac{\partial e_t}{\partial p_i}. \quad (6)$$

Providing the weights w_r are correctly chosen, each term H_{ij}^{-1} of the inverse matrix $\mathbf{H}^{-1}$ is related to the covariance between parameters i and j as follows (Cruickshank, 1965),

$$\mathrm{cov}(i,j) = H_{ij}^{-1}. \quad (7)$$

The diagonal elements of the inverse matrix thus give the variances of the corresponding parameters,

$$\sigma_i^2 = H_{ii}^{-1}. \quad (8)$$

The inverse matrix $\mathbf{H}^{-1}$ was computed by the matrix inversion routines *SPPFA* and *SPPDI* of *LINPACK* (Dongarra *et al.*, 1979), vectorized for the Convex C2 series computers. These routines implement the Cholesky decomposition method. Owing to limitations of real memory size, the inversions were performed in single precision arithmetic; comparisons performed in double precision with matrices of reduced size (*i.e.* half the number of elements) indicated that this caused no significant loss of precision.

The variances $\sigma^2(x)$, $\sigma^2(y)$ and $\sigma^2(z)$ of the atomic coordinates were obtained from the diagonal elements of $\mathbf{H}^{-1}$ using equation (8). From these, the standard deviations of the atomic positions $\langle|\Delta\mathbf{r}|^2\rangle^{1/2}$ were calculated using

$$\langle|\Delta\mathbf{r}|^2\rangle^{1/2} = [\sigma^2(x) + \sigma^2(y) + \sigma^2(z)]^{1/2}. \quad (9)$$

2.2. *Structures refined*

Two protein structures were used for the computation of the e.s.d.'s. Both proteins were crystallins, which are proteins found in the fibre cells of the eye lens. The first was γB-crystallin (previously called γII crystallin) whose structure was determined from X-ray diffraction data collected at the Daresbury synchrotron using photographic film (Wistow *et al.*, 1983; Najmudin *et al.*, 1993). The second structure used in this work was that of βB2-crystallin from X-ray diffraction data recorded on film at the Hamburg synchrotron (Bax *et al.*, 1990). The crystal data for the two proteins are listed in Table 1.

Table 1. *X-ray data for protein structures used for calculations of e.s.d.'s*

	γB-crystallin	βB2-crystallin
Data resolution (Å)	1.49	2.10
Space group	$P4_12_12$	$I222$
Protein molecules per asymmetric unit	1	1
Number of residues	175	177
Number of non-H protein atoms	1474	1466
Number of ordered solvent molecules	234	92
Number of reflections	26151	18583
Number of restraints	3821	3806
Number of observations (n)	29972	22389
Number of parameters (m)	6835	6259
$n - m$	23137	16130
Observation:parameter ratio	4.39	3.58
PDB code†	1gcs	2bb2

† Code associated with the coordinates deposited in the Protein Data Bank (Bernstein *et al.*, 1977).

Table 2. *Crystallographic R factors for γB- and βB2-crystallin for structure-factor calculations including only reflections within specified high-resolution cut-offs*

Figures in brackets are free R factors calculated from 5% of the reflection data which was omitted until the final stage of refinement.

Cut-off, $d_{\min}$ (Å)	γB-crystallin	βB2-crystallin
1.49	0.180 (0.204)	—
2.10	0.167	0.184 (0.200)
2.50	0.164	0.184
3.00	0.162	0.185

Each structure was rerefined using the weighting protocol described above and Table 2 shows the R factors of both structures which were calculated with different high-resolution cutoffs. No low-resolution cutoff was applied because solvent corrections to the calculated structure amplitudes were applied using Babinet's (1837) principle (see above).

3. Results

Fig. 1 shows a plot of the root-mean-square error $\langle|\Delta\mathbf{r}|^2\rangle^{1/2}$ against the isotropic atomic U for all atoms in the γB-crystallin structure, refined at 1.49 Å. It should be noted that the large majority of atoms have $U_{\rm iso}<0.25$ and $\langle|\Delta\mathbf{r}|^2\rangle^{1/2}<0.1$. Fig. 2 shows a similar plot where the standard deviations of individual atoms have been normalized according to their atomic numbers. This was accomplished by scaling to oxygen by multiplying $\langle|\Delta\mathbf{r}|^2\rangle^{1/2}$ by the atomic number of the atom divided by eight. Comparison with Fig. 1 shows that most points now lie close to or below the plotted curve. The cloud of outliers which fall below the curve arise from protein atoms which are subject to strong geometrical restraints which result in relatively lower $\langle|\Delta\mathbf{r}|^2\rangle^{1/2}$ values. Fig. 3 shows similar plots where the protein atoms have been

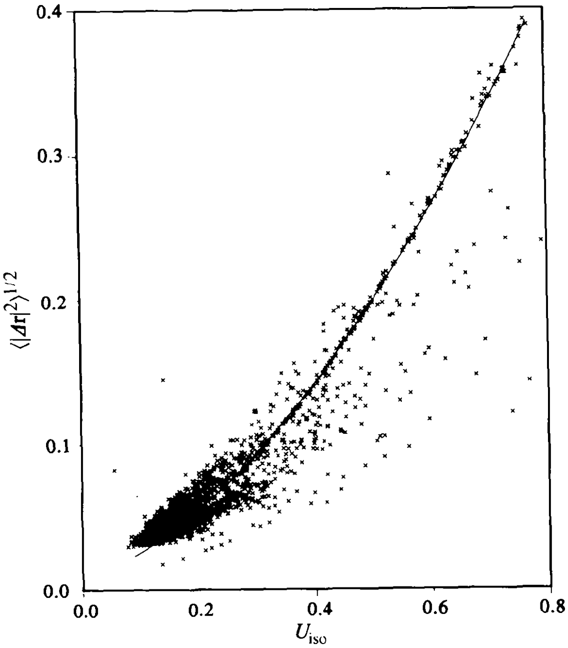

Fig. 1. The estimated $\langle|\Delta\mathbf{r}|^2\rangle^{1/2}$ (Å) for atoms in the γB-crystallin structure, refined at 1.49 Å, plotted against the isotropic atomic U value (Å^2). Atoms with U values greater than 0.8 Å^2 are not shown. The plotted curve has been fitted to the data points for the water atoms only. The curve has the form $\langle|\Delta\mathbf{r}|^2\rangle^{1/2} = a(U_{\rm iso} + b)^c$. The a, b and c parameters for the fitted curve are listed in Table 5 for 1.49 Å.

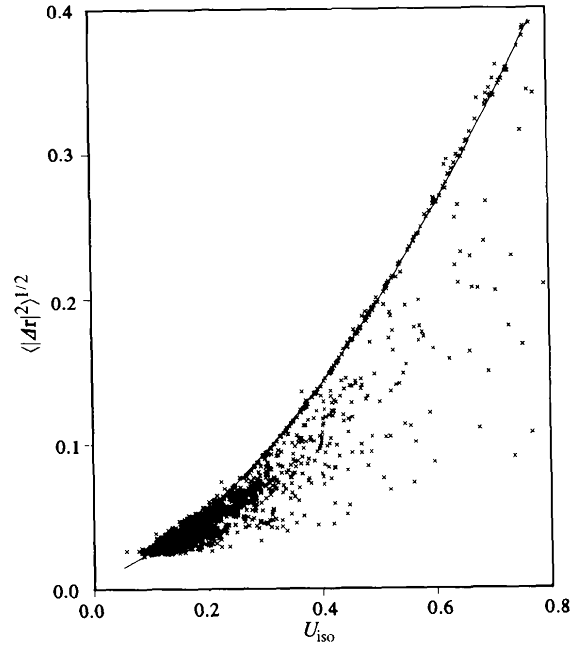

Fig. 2. Plot of $\langle|\Delta\mathbf{r}|^2\rangle^{1/2}$ (Å), normalized by atomic number, as a function of the isotropic atomic U value (Å^2) for the 1.49 Å refined structure of γB-crystallin. The plotted curve has been fitted to the data points for the water atoms only. The curve has the form $\langle|\Delta\mathbf{r}|^2\rangle^{1/2} = a(U_{\rm iso} + b)^c$. The curve is as in Fig. 1.

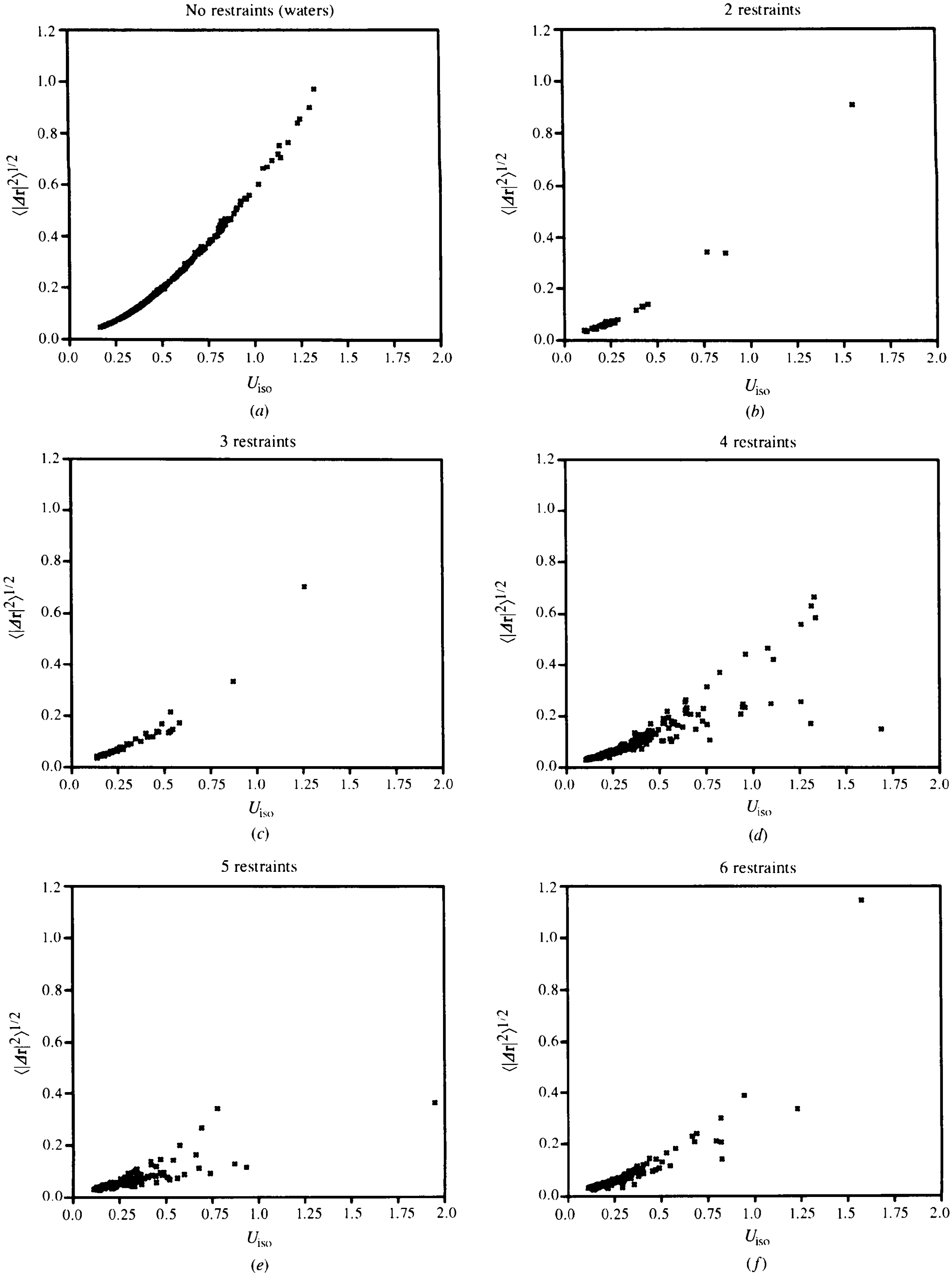

Fig. 3. Effects of the number of refinement restraints (distance as well as planarity restraints) on the atomic $\langle|\Delta\mathbf{r}|^2\rangle^{1/2}$ values. Each figure shows the normalized atomic $\langle|\Delta\mathbf{r}|^2\rangle^{1/2}$ values (Å) for the 1.49 Å refined structure of γB-crystallin (as in Fig. 2) as a function of the atom's isotropic *U* value (Å^2). The plots show: (*a*) all atoms with no restraints (*i.e.* water molecules); (*b*) atoms with two restraints; (*c*) three restraints; (*d*) four restraints; (*e*) five restraints; and (*f*) six restraints.

subdivided according to the number of restraints to which the atom is subject. A larger number of restraints results in relatively greater precision but also results in a less well defined relationship between $\langle|\Delta\mathbf{r}|^2\rangle^{1/2}$ and U. This is due to the differing effects of the different combinations of bond, bond angle and planar restraints.

Fig. 4 illustrates the effect of resolution on these all-atom plots from the two crystallin structures. The points are plotted for a truncated resolution of 2.10 Å in each case. The lines have been fitted to the water data assuming a simple relationship of the form $\langle|\Delta\mathbf{r}|^2\rangle^{1/2} = a(U_{\rm iso} + b)^c$. Average errors in atomic positions for groups of atoms $\overline{\Delta\mathbf{r}}$ have been estimated from the refinements at their highest resolutions and are displayed in Table 3 together with the associated overall temperature factor U_o for each group. Simple averaging of thermal parameters is sensitive to outlying values which can be very large and, therefore, poorly defined by the diffraction data. Such outliers are not uncommon in the side chains of protein molecules and can be seen in Fig. 4. The same considerations apply to averaging errors. Average errors calculated by the methods of Luzzati (1952) or Read (1986) are not sensitive to outliers. For comparison with our values a statistically robust averaging technique is required. Such a technique has been used in this paper and is described in *Appendix A*.

Table 3. *U_o values (Å^2) and average coordinate errors $\overline{\Delta\mathbf{r}}$ in atomic positions (Å) calculated by the method described in Appendix A*

	γB-crystallin ($d_{\rm min}$ 1.49 Å)		βB2-crystallin ($d_{\rm min}$ 2.10 Å)	
	U_o	$\overline{\Delta\mathbf{r}}$	U_o	$\overline{\Delta\mathbf{r}}$
Main-chain atoms	0.16	0.05	0.37	0.08
Side-chain atoms	0.22	0.14	0.44	0.20
All protein atoms	0.19	0.10	0.40	0.15
Water O atoms	0.45	0.27	0.65	0.35
All atoms	0.20	0.13	0.41	0.17
All atoms (Luzzati, 1952)		0.16		0.21
All atoms (Read, 1986)		0.12		0.17

4. Conclusions

4.1. *Mean-square deviations of the calculated restraints*

Fixing the scale of the structure-amplitude weights so that the residual minimized equals its statistical expec-

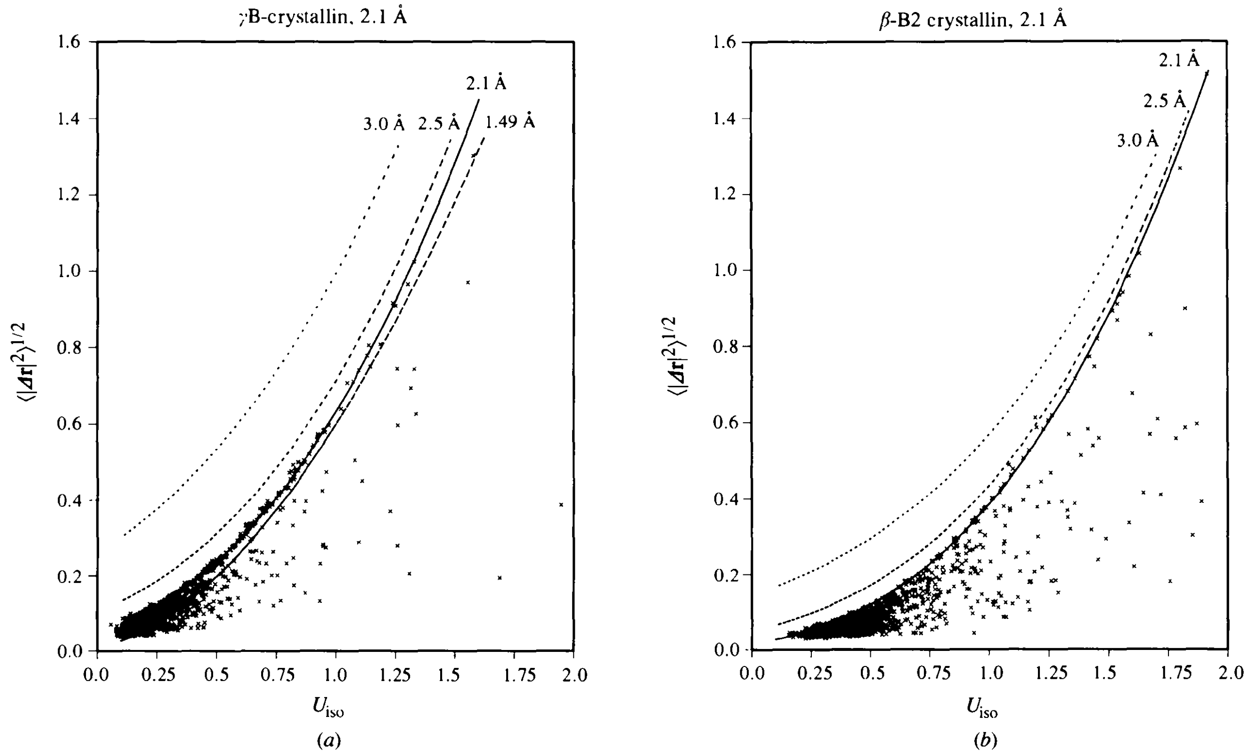

Fig. 4. The estimated $\langle|\Delta\mathbf{r}|^2\rangle^{1/2}$ (Å) as a function of the isotropic atomic U value (Å^2) for (*a*) γB-crystallin, calculated using data to 2.10 Å, and (*b*) βB2-crystallin refined to 2.10 Å. The $\langle|\Delta\mathbf{r}|^2\rangle^{1/2}$ values in both cases have been normalized to take atomic number into account. The curves on the plots show the best-fit lines for the water atoms. The lines have been fitted assuming the relationship $\langle|\Delta\mathbf{r}|^2\rangle^{1/2} = a(U_{\rm iso} + b)^c$. The a, b and c parameters for the fitted curves in the plots are listed in Table 5. The different curves on each plot give the results for e.s.d.'s calculated using data sets truncated to different resolutions: for γB-crystallin these are 1.49 Å, the resolution at which the structure was refined (also shown in Fig. 2), and three truncated data sets corresponding to 2.10, 2.50 and 3.00 Å; for βB2-crystallin the curves correspond to 2.10 Å (the resolution at which the structure was refined), 2.50 and 3.00 Å.

tation value, $n - m$, results in mean-square deviations of the calculated restraints from their target values which are smaller than the corresponding deviations observed by Engh & Huber (1991) in small-molecule structures and used for weighting the restraints in this work. The relevant mean-square deviations are displayed in Table 4 and show that the root-mean-square deviations of all distances at the convergence of the refinements are mostly less than two thirds of the Engh and Huber values. It is shown in *Appendix B* that the expected value of the calculated distance variance is

$$\left\langle \sum_{s}^{N_{\rm dist}} w_s[d_{\rm target}(s) - d_{\rm calc}(s)]^2 \right\rangle = N_{\rm dist} - {\rm tr}(\mathbf{H}^{-1}\mathbf{H}_{\rm dist}), \qquad (10)$$

where $\mathbf{H}_{\rm dist}$ is the normal matrix formed with distance terms only, w_s is the reciprocal of the Engh and Huber variance and tr denotes the trace operation. Evaluation of both sides of this equation was undertaken for βB2-crystallin at 2.1 Å resolution although calculation of the trace is only approximate since $\mathbf{H}$ is not a full matrix with respect to the temperature factors. The results are

$$\sum_{s}^{N_{\rm dist}} w_s[d_{\rm target}(s) - d_{\rm calc}(s)]^2 = 834$$

and

$$N_{\rm dist} - {\rm tr}(\mathbf{H}^{-1}\mathbf{H}_{\rm dist}) = 3546 - 2630 = 916.$$

It should also be noted that the minimum variance structure will have a higher R than a structure refined so as to get mean-square deviations from ideal geometry equal to Engh and Huber variances. We suspect that many crystallographers just aim at the lowest R factor which is consistent with results from a validation program such as *PROCHECK* (Laskowski *et al.*, 1993) that are thought to be acceptable to a referee!

4.2. *Average errors in atomic coordinates*

The average errors in atomic positions calculated from the refinements and shown in Table 3 take into account the random errors in the data and also such errors in the model that produce random fluctuations in the residuals but not any silent errors in the model. As expected for a restrained refinement, the errors in the protein atoms depend strongly on the number of geometrical restraints associated with the particular atom. These errors may be highly anisotropic.

Comparison of average errors in protein coordinates will always be sensitive to the averaging technique due to the very skew distribution of error among the atoms. Using the robust averaging technique described in *Appendix A*, our all-atom error estimates in Table 3 are lower than those produced by the method of Luzzati but in good agreement with the method of Read.

Table 4. *Root-mean square deviations from target values at the end of refinement compared with literature standard deviations derived from Engh & Huber (1991) (Å)*

	γB- crystallin	βB2- crystallin	Standard deviation
$d_{\rm min}$	1.49	2.10	
Bond distances (1–2)	0.010	0.009	0.022
Angle distances (1–3)	0.025	0.022	0.037
All distances	0.020	0.018	0.032

Table 5. *Parameters for the various curves in Fig. 4, fitted to the data for the water atoms using curves of the form* $\langle|\Delta\mathbf{r}|^2\rangle^{1/2} = a(U_{iso} + b)^c$

Structure	Resolution cut-off (Å)	Parameters a	b	c
γB-crystallin	1.49	0.513	0.090	1.80
	2.10	0.190	0.591	2.59
	2.50	0.0651	1.150	3.13
	3.00	0.0068	2.478	4.00
βB2-crystallin	2.10	0.127	0.478	2.84
	2.50	0.0194	1.282	3.77
	3.00	1.11×10^{-6}	4.909	7.40

Table 5 shows the coefficients used in the polynomials that generated the curves in Fig. 4 which describe the relationship between the standard deviations of the water positions and their temperature factors. It can be seen that the power of the temperature factor in these formulae increases with resolution. This is because at lower resolution, waters with a high temperature factor are more poorly defined. Cruickshank (1996) suggested a similar expression for errors which was quadratic in U and which gives qualitative agreement with our results at about 1.6 Å resolution. Although the functional form of the relationship between the coordinate standard deviation and temperature factor is simple, values of the constants in the formula depend on the quality of the raw X-ray data. This makes the derivation of a formula linking standard deviation and temperature factor a difficult task which we shall attempt in future work.

APPENDIX *A*
The 'average' isotropic thermal parameter and 'average' coordinate error

Simple averaging of thermal parameters is sensitive to outlying values which can be very large and therefore poorly defined by the diffraction data. Such outliers are not uncommon in the side chains of protein molecules. The same considerations apply to the averaging of errors. What is needed is a statistically robust averaging technique. The method used in this paper is based on averaging the exponential temperature-factor expressions, rather than averaging the U values themselves. Blessing *et al.* (1996) have described a method for averaging these expressions but their method assumes that the U values are normally distributed. Since the distribution may be skew in a macromolecular crystal, we use a method that avoids this assumption.

Assume we have two identical models, except that in one the atoms have individual isotropic thermal parameters U_j, and in the other the atoms have the same overall $U = U_o$. Then define the 'average' U of the first model to be equal to U_o when $\sum |\mathbf{F}_h|^2$ is the same for both models. This is not necessarily the optimum definition, but it does have the advantage of leading to a reasonably tractable solution.

The 'average' coordinate error is obtained in the same way by treating it as equivalent to an isotropic thermal parameter with a value equal to one third the mean-square coordinate error (Read, 1986),

$$U' = \langle|\Delta\mathbf{r}|^2\rangle/3.$$

Assuming Wilson's statistics (randomly distributed atoms),

$$\overline{|\mathbf{F}_h|}^2_{\mathbf{s}} = T_o^2(\mathbf{s})\sum_j f_j^2(\mathbf{s}) = \sum_j f_j^2(\mathbf{s})T_j^2(\mathbf{s}),$$

where the overbar indicates arithmetic mean throughout, $\mathbf{s}$ is the reciprocal lattice vector, the temperature factor is given by $T_o(\mathbf{s}) = \exp(-2\pi^2|\mathbf{s}|^2U_o)$, and $T_j(\mathbf{s}) = \exp(-2\pi^2|\mathbf{s}|^2U_j)$.

Now, assuming equal atoms, replacing the sum over atoms with an arithmetic mean, and integrating over reciprocal lattice points within a sphere of radius $|\mathbf{s}|_{\max}$,

$$\int_0^{|\mathbf{s}|_{\max}} T_0^2(\mathbf{s})|\mathbf{s}|^2\mathrm{d}|\mathbf{s}| = \int_0^{|\mathbf{s}|_{\max}} \overline{T_j^2(\mathbf{s})|\mathbf{s}|^2\mathrm{d}|\mathbf{s}|}.$$

Substituting for $T_o(\mathbf{s})$ and $T_j(\mathbf{s})$

$$\int_0^{|\mathbf{s}|_{\max}} \exp(-4\pi^2|\mathbf{s}|^2U_o)|\mathbf{s}|^2\mathrm{d}|\mathbf{s}| = \int_0^{|\mathbf{s}|_{\max}} \overline{\exp(-4\pi^2|\mathbf{s}|^2U_j)|\mathbf{s}|^2\mathrm{d}|\mathbf{s}|}.$$

Therefore,

$$x_o^{-3}\int_0^{x_o}\exp(-t^2/2)t^2\mathrm{d}t = \overline{x_j^{-3}\int_0^{x_j}\exp(-t^2/2)t^2\mathrm{d}t}, \quad (11)$$

where $x_o = 2^{3/2}\pi U_o^{1/2}|\mathbf{s}|_{\max}$ and $x_j = 2^{3/2}\pi U_j^{1/2}|\mathbf{s}|_{\max}$.

The integrals on both sides of equation (11) can be evaluated as

$$\int_0^x \exp(-t^2/2)t^2\mathrm{d}t = \int_0^x \exp(-t^2/2)\mathrm{d}t - x\exp(-x^2/2) = (\pi/2)^{1/2}\Phi(x) - x\exp(-x^2/2),$$

where Φ is the 'error integral'.

Using this result and rearranging (11) we obtain

$$\Phi(x_o) - (2/\pi)^{1/2}x_o\exp(-x_o^2/2) - Ax_o^3 = 0, \quad (12)$$

where

$$A = \overline{x_j^{-3}[\Phi(x_j) - (2/\pi)^{1/2}x_j\exp(-x_j^2/2)]}.$$

The value of A is computed for the set of atoms whose U's are to be averaged, and then (12) solved for x_o, and hence U_o, by Newton's method starting from an initial guess for x_o. To solve (12), where the left-hand side is written as $y(x_o)$, shifts of magnitude $-y(x_o)/y'(x_o)$ are applied iteratively to x_o until convergence, where the derivative of $y(x_o)$ is

$$y'(x_o) = (2/\pi)^{1/2}x_o^2\exp(-x_o^2/2) - 3Ax_o^2.$$

A good initial guess for x_o is

$$x_o^0 = A^{1/3} - 2(\pi/A)^2.$$

Finally

$$U_o = x_o^2/(8\pi^2|\mathbf{s}|^2_{\max}).$$

APPENDIX *B*
The expected value of least-squares residuals

In this *Appendix* we first show that second moment matrix of least-squares residuals is equal to the difference between the variance–covariance matrix (VCM) of the observations (structure amplitudes and distances) and the VCM of the corresponding quantities calculated from the parameter estimates at the convergence of a refinement. Subtraces of the second moment matrix of the residuals yield the expected values of the sum of subsets of residuals, including the case of the expected value of the distance residuals which is shown to be smaller than the variance of the observed distances.

We define the following column matrices:

$\mathbf{f}$ the observations (structure amplitudes and target distances);

$\hat{\mathbf{f}}$ the least-squares estimates of $\mathbf{f}$ calculated at the convergence of the refinement;

$\hat{\mathbf{x}}$ the least-squares estimates of the parameters calculated from $\hat{\mathbf{f}}$;

$\boldsymbol{\Delta} = \mathbf{f} - \hat{\mathbf{f}}$ the least-squares residuals.

We further define the following rectangular matrices:

$\mathbf{A}$ the least-squares design matrix of order $n \times m$ where n is the number of observations and m is the number of parameters;

$\mathbf{W}$ the $n \times n$ symmetric weight matrix and $\mathbf{W}^{-1} = \langle(\mathbf{f} - \langle\mathbf{f}\rangle)(\mathbf{f} - \langle\mathbf{f}\rangle)^T\rangle$ is the VCM of the observations where $\langle\mathbf{f}\rangle$ is the expected value of $\mathbf{f}$;

$\mathbf{H}$ the $m \times m$ normal matrix given by $\mathbf{A}^T\mathbf{W}\mathbf{A}$;

$\mathbf{R}$ an $n \times n$ idempotent matrix ($\mathbf{R}^2 = \mathbf{R}$) given by $\mathbf{A}\mathbf{H}^{-1}\mathbf{A}^T\mathbf{W}$.

Given these definitions the normal equations at convergence can be written as

$$\begin{aligned}\mathbf{0} &= (\mathbf{A}^T\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{W}\boldsymbol{\Delta}\\ &= \mathbf{H}^{-1}\mathbf{T}^T\mathbf{W}\boldsymbol{\Delta}.\end{aligned}$$

If the errors in the observations are not too large then truncated Taylor expansions may be written about the expected values of the parameter vector $\langle\mathbf{x}\rangle$ and the observation vector $\langle\mathbf{f}\rangle$.

$$\begin{aligned}\hat{\mathbf{x}} - \langle\mathbf{x}\rangle &= \mathbf{H}^{-1}\mathbf{A}^T\mathbf{W}(\mathbf{f} - \langle\mathbf{f}\rangle)\\ \hat{\mathbf{f}} - \langle\mathbf{f}\rangle &= \mathbf{A}(\hat{\mathbf{x}} - \langle\mathbf{x}\rangle).\end{aligned}$$

Hence,

$$\begin{aligned}\hat{\mathbf{f}} - \langle\mathbf{f}\rangle &= \mathbf{A}\mathbf{H}^{-1}\mathbf{A}^T\mathbf{W}(\mathbf{f} - \langle\mathbf{f}\rangle)\\ &= \mathbf{R}(\mathbf{f} - \langle\mathbf{f}\rangle).\end{aligned} \qquad (13)$$

Using the definition of $\mathbf{R}$, the second moment matrix of the residuals $\mathbf{D} = \langle\boldsymbol{\Delta}\boldsymbol{\Delta}^T\rangle$ is given by

$$\begin{aligned}\mathbf{D} &= (\mathbf{I}_n - \mathbf{R})\mathbf{W}^{-1}(\mathbf{I}_n - \mathbf{R})^T\\ &= \mathbf{W}^{-1} + \mathbf{R}\mathbf{W}^{-1}\mathbf{R}^T - 2\mathbf{R}\mathbf{W}^{-1}\\ &= \mathbf{W}^{-1} - \mathbf{R}\mathbf{W}^{-1},\end{aligned} \qquad (14)$$

where $\mathbf{I}_n$ is a unit matrix of order n. From (13) the VCM of $\hat{\mathbf{f}}$ is

$$\langle(\hat{\mathbf{f}} - \langle\mathbf{f}\rangle)(\hat{\mathbf{f}} - \langle\mathbf{f}\rangle)^T\rangle = \mathbf{R}\mathbf{W}^{-1},$$

and thus from (14) we see that $\mathbf{D}$ is equal to the difference between the VCM of the observations and the VCM of the corresponding quantities calculated from the refined parameters.

Equation (14) can be rewritten as

$$\mathbf{D}\mathbf{W} = \mathbf{I} - \mathbf{R}. \qquad (15)$$

If $\mathbf{A}_p$ is the matrix containing p rows of $\mathbf{A}$, and $\mathbf{D}_p\mathbf{R}_p$ and $\mathbf{W}_p$ are the corresponding $p \times p$ diagonal submatrices of $\mathbf{D}$, $\mathbf{R}$ and $\mathbf{W}$, respectively, then

$$\mathbf{D}_p\mathbf{W}_p = \mathbf{I}_p - \mathbf{R}_p.$$

Taking the trace of both sides

$$\begin{aligned}\mathrm{tr}(\mathbf{D}_p\mathbf{W}_p) &= p - \mathrm{tr}(\mathbf{A}_p\mathbf{H}^{-1}\mathbf{A}_p^T\mathbf{W}_p)\\ &= p - \mathrm{tr}(\mathbf{H}^{-1}\mathbf{A}_p^T\mathbf{W}_p\mathbf{A}_p)\\ &= p - \mathrm{tr}(\mathbf{H}^{-1}\mathbf{H}_p)\end{aligned} \qquad (16)$$

where $\mathbf{H}_p = \mathbf{A}_p^T\mathbf{W}_p\mathbf{A}_p$.

If the p rows of $\mathbf{A}_p$ correspond to the distance restraints and we assume that the weight matrix is diagonal, then (16) can be written as

$$\left\langle \sum_s^{N_{\rm dist}} w_s[d_{\rm target}(s) - d_{\rm calc}(s)]^2 \right\rangle = N_{\rm dist} - \mathrm{tr}(\mathbf{H}^{-1}\mathbf{H}_{\rm dist})$$

which is (10). From (15) the expected value of the ith squared distance residual is given by

$$\langle w_i[d_{\rm target}(i) - d_{\rm calc}(i)]^2\rangle = 1 - R_{ii},$$

where R_{ii} is the relevant diagonal element of $\mathbf{R}$.

Thus

$$\begin{aligned}\langle[d_{\rm target}(i) - d_{\rm calc}(i)]^2\rangle &= \sigma_E^2(1 - R_{ii})\\ &= \sigma_E^2 - \mathbf{a}_i^T\mathbf{H}^{-1}\mathbf{a}_i\end{aligned}$$

where σ_E^2 is the variance given by Engh & Huber (1991) and $\mathbf{a}_i$ is the relevant row of $\mathbf{A}$. The expected mean-square distance residual is less than σ_E^2 because $\mathbf{a}_i^T\mathbf{H}^{-1}\mathbf{a}_i$ is a positive definite quadratic form.

A well known special case of (16) occurs when $p = n$. In this case (16) yields the expected value of the total residual $\langle M\rangle$.

$$\begin{aligned}\langle M\rangle &= p - \mathrm{tr}(\mathbf{H}^{-1}\mathbf{H}_p)\\ &= n - \mathrm{tr}(\mathbf{I}_m)\\ &= n - m.\end{aligned}$$

This result was used in the present work for determining the scale of the structure-amplitude weights.

The authors would like to thank Professor D. W. J. Cruickshank FRS for his helpful comments on this paper.

References

Babinet, J. (1837). *C. R. Acad. Sci. Paris*, **4**, 638–648.

Bax, B., Lapatto, R., Nalini, V., Driessen, H., Lindley, P. F., Mahadevan, D., Blundell, T. L. & Slingsby, C. (1990). *Nature (London)*, **347**, 776–780.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Blessing, R. H., Guo, D. Y. & Langs, D. A. (1996). *Acta Cryst.* D**52**, 257–266.

Carson, M., Buckner, T. W., Yang, Z., Narayana, V. L. & Bugg, C. E. (1994). *Acta Cryst.* D**50**, 900–909.

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. (1965). *Errors in least-squares methods*, in *Computing Methods in Crystallography*, edited by J. S. Rollett, pp. 112–116. Oxford: Pergamon.

Cruickshank, D. W. J. (1996). *Protein precision re-examined: Luzzati plots do not estimate final errors*, in *Proceedings of the CCP4 Study Weekend*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey. Warrington: Daresbury Laboratory.

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

Dongarra, J. J., Moler, C. B., Bunch, J. R. & Stewart, G. W. (1979). *Linpack Users' Guide*, Siam, Philadelphia.

Driessen, H., Haneef, M. I. J., Harris, G. W., Howlin, B., Khan, G. & Moss, D. S. (1989). *J. Appl. Cryst.* **22**, 510–516.

Elango, N. & Parthasarathy, S. (1990). *Acta Cryst.* A**46**, 495–502.

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.

Hamilton, W. C. (1964). *Statistics in Physical Science*. New York: Ronald Press.

Haneef, I., Moss, D. S., Stanford, M. J. & Borkakoti, N. (1985). *Acta Cryst.* A**41**, 426–433.

Holland, D. R., Clancy, L. L., Muchmore, S. W., Ryde, T. J., Einspahr, H. M., Finzel, B. C., Heinrikson, R. L. & Watenpaugh, K. D. (1990). *J. Biol. Chem.* **265**, 17649–17656.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

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

Moss, D. S. & Morffew, A. J. (1982). *Comput. Chem.* **6**(1), 1–3.

Murshudov, G. & Dodson, E. J. (1997). *Simplified error estimation a la Cruickshank in macromolecular crystallography*, in *CCP4 Newsletter*, January 1997, edited by S. Bailey. Warrington: Daresbury Laboratory.

Najmudin, S., Nalini, V., Driessen, H. P. C., Slingsby, C., Blundell, T. L., Moss, D. S. & Lindley, P. F. (1993). *Acta Cryst.* D**49**, 223–233.

Nielsen, K. (1977). *Acta Cryst.* A**33**, 1009–1010.

Ohlendorf, D. H. (1994). *Acta Cryst.* D**50**, 808–812.

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

Robinson, W. T. & Sheldrick, G. M. (1988). *Crystallographic Computing 4: Techniques and New Technologies*, edited by N. W. Isaacs & M. R. Taylor, pp. 366–377. IUCr/Oxford University Press.

Sevcik, J., Hill, C. P., Dauter, Z. & Wilson, K. S. (1993). *Acta Cryst.* D**49**, 257–271.

Sheldrick, G. M. (1976). *SHELX76. Program for Crystal Structure Determination*. University of Cambridge, England.

Sheldrick, G. M. (1985). *Crystallographic Computing 3: Data Collection, Structure Determination, Proteins, and Databases*, edited by G. M. Sheldrick, C. Krüger & R. Goddard, pp. 184–189. Oxford: Clarendon Press.

Sheldrick, G. M. (1986). *SHELX86. Program for Crystal Structure Determination*. University of Göttingen, Germany.

Stroud, R. M. & Fauman, E. B. (1995). *Protein Sci.* **4**, 2392–2404.

Wistow, G., Turnell, B., Summers, L., Slingsby, L., Moss, D., Miller, L., Lindley, P. & Blundell, T. (1983). *J. Mol. Biol.* **170**, 175–202.